

NLP+CSS 2024

**The Sixth Workshop on Natural Language Processing and
Computational Social Science**

Proceedings of the Workshop

June 21, 2024

©2024 published of the conference, generally Association for Computational Linguistics"

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-112-4

Introduction

Welcome to the Sixth Workshop on Natural Language Processing (NLP) and Computational Social Science (CSS)! This workshop continues a successful series, with many interdisciplinary contributions to make NLP techniques and insights standard practice in CSS research, as well as improve NLP through insights from the social sciences. We received 54 submissions and after a rigorous review process by our committee, we accepted 15 entries, 11 archival and 4 non-archival. We are excited to see a diverse range of topics and disciplines, and we hope to continue the tradition of fostering a dialogue between researchers in NLP and social science. Our workshop program also includes keynote talks by three outstanding scholars: Naoki Egami, a professor in the Department of Political Science at Columbia University, Maria Antoniak, a Young Investigator at the Allen Institute for AI, and Helena Gómez Adorno, a professor in the Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas (IIMAS) at the Universidad Nacional Autónoma de México (UNAM).

We would like to thank the Program Committee members who reviewed the papers this year. They did a heroic job providing some top-notch reviews in a short time frame, and particularly helping with last minute emergency reviews. We would also like to thank the workshop participants for the opportunities to connect (or reconnect) and learn from each other.

Dallas Card, Anjalie Field, Dirk Hovy, Katherine Keith (Co-Organizers)

Organizing Committee

Organizers

Dallas Card, University of Michigan
Anjalie Field, Johns Hopkins University
Dirk Hovy, Bocconi University
Katherine Keith, Williams College

Reviewers

Prabhat Agarwal, Computer Science Department, Stanford University
Jisun An, Indiana University
Dnyaneshwar Bhadane, Indian Institute of Information Technology Lucknow
Aoife Cahill, Dataminr
Sky CH-Wang, Columbia University
Hyundong Justin Cho, USC/ISI
Krish Chopra, Netaji Subhas University of Technology
Aron Culotta, Tulane University
Walter Daelemans, University of Antwerp
Aida Mostafazadeh Davani, Research, Google
Jonathan Dunn, University of Illinois Urbana-Champaign
Blanca Figueras, Universidad del País Vasco
Kathleen C. Fraser, National Research Council Canada
Lea Frermann, University of Melbourne
Nabeel Gillani, Northeastern University
Pranav Goel, Northeastern University
Graeme Hirst, University of Toronto
Alexander Hoyle, University of Maryland, College Park
Tiancheng Hu, University of Cambridge
Kristen Johnson, Michigan State University
Kenneth Joseph, State University of New York at Buffalo
David Jurgens, University of Michigan - Ann Arbor
Elma Kerz, Rheinisch Westfälische Technische Hochschule Aachen
Philipp Koehn, Johns Hopkins University
Haewoon Kwak, Indiana University
Oliver Li, Columbia University
Sijia Liu, Amazon
Li Lucy, Allen Institute for Artificial Intelligence
Ziqian Luo, Oracle
Tessa Masis, University of Massachusetts at Amherst
Diana Maynard, University of Sheffield
Julia Mendelsohn, University of Michigan - Ann Arbor
Rada Mihalcea, University of Michigan
Seyed Abolghasem Mirroshandel, University of Guilan
Kunihiro Miyazaki, Indiana University
Arjun Mukherjee, University of Houston
Nazia Nafis, The University of Sheffield
Bill Noble, Göteborg University
Pierre Nugues, Lund University

Brendan O'Connor, University of Massachusetts, Amherst
Sebastian Padó, University of Stuttgart, Universität Stuttgart
Ankit Pal, Saama AI Research Lab
Chanjun Park, Upstage
Vinodkumar Prabhakaran, Google
Akash Rawat, Indian Institute of Information Technology, Lucknow
Georg Rehm, Humboldt Universität Berlin
Manoel Horta Ribeiro, EPFL - EPF Lausanne
Matiss Rikters, National Institute of Advanced Industrial Science and Technology (AIST)
Ellen Riloff, University of Arizona
Anthony Rios, University of Texas at San Antonio
Frank Rudzicz, Dalhousie University
Djamé Calvo Seddah, Inria Paris (on teaching leave from Sorbonne Université)
Andreas Spitz, Universität Konstanz
Maximilian Spliethöver, Leibniz University Hannover
Vivek Srivastava, Tata Consultancy Services Limited, India
Maja Stahl, Leibniz Universität Hannover
Ian Stewart, Pacific Northwest National Laboratory
Marek Suppa, Comenius University in Bratislava
Manan Suri, Netaji Subhas Institute of Technology
Samia Touileb, University of Bergen
Vasudha Varadarajan, State University of New York, Stony Brook
Rob Voigt, Northwestern University
Ming Feng Wan, University of Toronto
Charles Welch, Rheinische Friedrich-Wilhelms Universität Bonn
Winston Wu, University of Hawaii at Hilo
Bernardo Nugroho Yahya, Hankuk University of Foreign Studies
Jinrui Yang, The University of Melbourne
François Yvon, Université Pierre et Marie Curie - Paris 6, Sorbonne Université - Faculté des Sciences (Paris VI)
Yongjun Zhang, State University of New York at Stony Brook
Mian Zhong, Swiss Federal Institute of Technology

Table of Contents

<i>Detecting Perspective-Getting in Wikipedia Discussions</i> Evgeny Vasilets, Tijs A. Van Den Broek, Anna Wegmann, David Abadi and Dong Nguyen	1
<i>Connecting the Dots in News Analysis: Bridging the Cross-Disciplinary Disparities in Media Bias and Framing</i> Gisela Vallejo, Timothy Baldwin and Lea Frermann	16
<i>The Crime of Being Poor: Associations between Crime and Poverty on Social Media in Eight Countries</i> Georgina Curto, Svetlana Kiritchenko, Kathleen C. Fraser and Isar Nejadgholi	32
<i>Discovering Implicit Meanings of Cultural Motifs from Text</i> Anurag Acharya, Diego Estrada, Shreeja Dahal, W. Victor H. Yarlott, Diana Gomez and Mark Finlayson	46
<i>Can Large Language Models (or Humans) Disentangle Text?</i> Nicolas Audinet De Pieuchon, Adel Daoud, Connor Thomas Jerzak, Moa Johansson and Richard Johansson	57
<i>Retrieval Augmented Generation of Subjective Explanations for Socioeconomic Scenarios</i> Razvan-Gabriel Dumitru, Maria Alexeeva, Keith Alcock, Nargiza Ludgate, Cheonkam Jeong, Zara Fatima Abdurahaman, Prateek Puri, Brian Kirchoff, Santadarshan Sadhu and Mihai Surdeanu	68
<i>Where on Earth Do Users Say They Are?: Geo-Entity Linking for Noisy Multilingual User Input</i> Tessa Masis and Brendan O'Connor	86
<i>News Deja Vu: Connecting Past and Present with Semantic Search</i> Brevin Franklin, Emily Silcock, Abhishek Arora, Tom Bryan and Melissa Dell	99
<i>Knowledge Distillation in Automated Annotation: Supervised Text Classification with LLM-Generated Training Labels</i> Nicholas J Pangakis and Sam Wolken	113
<i>Clustering Document Parts: Detecting and Characterizing Influence Campaigns from Documents</i> Zhengxiang Wang and Owen Rambow	132
<i>A First Step towards Measuring Interdisciplinary Engagement in Scientific Publications: A Case Study on NLP + CSS Research</i> Alexandria Leto, Shamik Roy, Alexander Hoyle, Daniel Acuna and Maria Leonor Pacheco . .	144

Detecting Perspective-Getting in Wikipedia Discussions

Evgeny Vasilets[♡] Tijds Van den Broek[♣] Anna Wegmann[♠]

David Abadi[♡] Dong Nguyen[♠]

[♡] University of Amsterdam, The Netherlands

[♣] VU Amsterdam, The Netherlands

[♠] Utrecht University, The Netherlands

t.a.vanden.broek@vu.nl, a.m.wegmann@uu.nl

Abstract

Perspective-getting (i.e., the effort to obtain information about the other person’s perspective) can lead to more accurate interpersonal understanding. In this paper, we develop an approach to measure perspective-getting and apply it to English Wikipedia discussions. First, we develop a codebook based on perspective-getting theory to operationalize perspective-getting into two categories: asking questions about and attending the other’s perspective. Second, we use the codebook to annotate perspective-getting in Wikipedia discussion pages. Third, we fine-tune a RoBERTa model that achieves an average F-1 score of 0.76 on the two perspective-getting categories. Last, we test whether perspective-getting is associated with discussion outcomes. Perspective-getting was not higher in non-escalated discussions. However, discussions starting with a post attending the other’s perspective are followed by responses that are more likely to also attend the other’s perspective. Future research may use our model to study the influence of perspective-getting on the dynamics and outcomes of online discussions.

1 Introduction

Understanding what other individuals think or feel is important for effective interpersonal behavior, such as building and maintaining social relationships (Hughes and Leekam, 2004; Keysar et al., 2003), effective customer service (Axtell et al., 2007), or reaching agreements in negotiations (Galinsky et al., 2008; Trötschel et al., 2011). Individuals take another individual’s perspective (i.e., engage in *perspective-taking*) by putting themselves in the same “mental shoes” as the other (Galinsky et al., 2008) and by recognizing that this other person’s perspective differs from their ego-centric perspective (Epley et al., 2004a). Scholars have shown that imagining someone else’s

perspective and incorporating this alternative viewpoint in communication is a notoriously difficult task (Eyal et al., 2018; Epley et al., 2004a; Damen et al., 2020, 2019). Individuals may take their own perspective as an anchor for assessing other’s perspective, lack motivations to seek for information about other’s perspective, or consider perspective-taking as requiring too much mental effort (Epley et al., 2004a,b). Consequently, individuals often fail to accurately infer the perspective of others (Realo et al., 2003; Eyal et al., 2018).

In contrast to the cognitive process of imagining an other’s perspective through perspective-taking, the strategy of **perspective-getting** results in a more accurate assessment of the other’s perspective (Bruneau and Saxe, 2012; Eyal et al., 2018; Kalla and Broockman, 2023). Perspective-getting refers to directly asking about and listening to information related to another person’s thoughts, feelings, beliefs, or preferences (Damen et al., 2021; Eyal et al., 2018; Kalla and Broockman, 2023) instead of simply imagining the other’s perspective. In contrast to perspective-taking, perspective-getting can likely be more accurately annotated and detected as it is about detecting visible behavior as opposed to cognitive states from text. Perspective-getting has been operationalized as whether the interlocutor asks a target interlocutor about their perspective (Eyal et al., 2018; Damen et al., 2021) and as whether interlocutors summarize the other’s perspective (Kalla and Broockman, 2023).

Previous work manipulated perspective-getting via participant instructions in experimental settings (Bruneau and Saxe, 2012; Eyal et al., 2018; Damen et al., 2021; Kalla and Broockman, 2023). To the best of our knowledge, perspective-getting was not studied outside lab settings and has not been investigated computationally. We study perspective-getting in a natural online discussion settings on Wikipedia and build a classifier for computational analysis.

	Shortened Example
Attending the Other’s Perspective (AOP)	<p>✓ Message: MOS:INTRO states for the LEDE explicit “Do not hint at startling facts without describing them”. I suppose that the “War beginning” is such a startling facts , @DS: why do you refuse to give the reader the information about issues 1-4 in the lede?.</p> <p>Reply: To state when the war started is not a “startling fact”. I see no reason to mention in the lede that the Chilean troops capturing Antofagasta got a warm welcome, it is quite simply an undue bit of trivia for the lede.</p>
	<p>✗ Message: If the statement is not in serious dispute (i.e. in high-quality sources) then it’s a fact we can just assert. Failing to do so would not be neutral. </p> <p>Reply: We’ve gotten circular at this point, need input from others.</p>
Asking for Clarification (AFC)	<p>✓ Message: This article barely reads like an article. Opinions are stated as facts, speculations are presented as conclusions, etc., and numerous questions are omitted.</p> <p>Reply: It would help if you were more specific.</p>
	<p>✗ Message: Isabel Allende is the Spanish speaking equivalent to English author E.L. James. She is the author of the fifty shades of grey books. I’m curious, would you use a biography of E.L. James to validate an assertion on a war history article? I’m going out on a limb here but it appears your clearly out of you depth when comes to Latin american history articles, the fact your clearly using a quote finder and not being scrupulous with the quote and the source validates my point.</p> <p>Reply: You do not get to dismiss a source just because you do not like it, not a single piece of text you have written actually disputes the sources I have supplied, cheers.</p>

Table 1: *Perspective-getting*. For our perspective-getting categories AOP and AFC, we display two example messages, one displaying (✓) and one not displaying (✗) the category. We also display the previous message (Message) that the considered message replies to (Reply). Annotators see both messages during annotation, RoBERTa only sees the Reply. Perspective-getting refers to AFC about the other’s perspective (Damen et al., 2021; Eyal et al., 2018) and listening to the other’s perspective — which was operationalized with summarization by Kalla and Broockman (2023) and Bruneau and Saxe (2012). Our AOP category encompasses summarization.

Approach and contributions In this study, we aim to measure perspective-getting in messages in online discussions.

First, we develop a codebook to operationalize perspective-getting into two categories: Attending an Other’s Perspective (AOP) and Asking For Clarification (AFC). Our codebook is grounded in previous literature on perspective-getting (Kalla and Broockman, 2023; Eyal et al., 2018; Damen et al., 2021) and contains examples from Wikipedia discussions. We specifically focus on discussions between editors on Wikipedia, during which editors have to reach a consensus on a difference in point of view (Black et al., 2011; De Kock and Vlachos, 2021, 2022). We expect this to be a relevant setting to study perspective-getting as, to come to a joint resolution, it should be important to understand each other’s perspective.

Second, we annotate perspective-getting in Wikipedia discussions and then fine-tune a RoBERTa model (Liu et al., 2019), which achieves a macro average F1 score of 0.76.

Third, we test whether perspective-getting scores are associated with (i) the escalation of Wikipedia discussions to a moderator and (ii) the development of perspective-getting in Wikipedia discussions. Perspective-getting was not higher in non-

escalated discussions. However, discussions starting with a post attending the other’s perspective (AOP) are more often followed by responses that show AOP as well. The codebook, dataset and code are included in our project’s [OSF repository](#).

2 Background

Perspective-getting (Eyal et al., 2018; Damen et al., 2021; Kalla and Broockman, 2023) has been a more recent term to differentiate the cognitive process of perspective-taking from strategies that aim at “getting” the perspective from another person (e.g., directly asking about their perspective). This understanding of perspective-getting has previously been and sometimes still is incorporated in some definitions of perspective-taking (Bruneau and Saxe, 2012; Kalla and Broockman, 2020, 2023). In this work, we focus on perspective-getting but we include insights from “perspective-taking” studies if they operationalize it similar to our understanding of perspective-getting as defined in Section 3.

Perspective-getting has not yet been measured using NLP methods. But our two categories for perspective-getting (AOP and AFC, see also Table 1 and Section 3) are related to various research topics in NLP. Attending the Other’s Perspective (AOP) can manifest in various ways, including

quoting, paraphrasing or summarizing the other’s message. In NLP, various studies have investigated quotation detection (Pareti et al., 2013), paraphrase detection (Dolan and Brockett, 2005; He et al., 2020; Kanerva et al., 2023), and summarization (El-Kassas et al., 2021; Tuggener et al., 2021). However, AOP can not be fully represented by any of them and even a combination would face new challenges in the dialog setting.

Asking for Clarification (AFC) is related to detecting questions, which can be challenging in online messages as, for example, question marks can be absent or questions might be phrased as information requests (Wang and Chua, 2010). Transformer-based architectures have been used with moderate success to detect questions in informal online texts (Saha et al., 2020, 2021; Kalouli et al., 2021). For clarification questions specifically, there exist corpora (Rao and Daumé III, 2018; Xu et al., 2019) and generative models (Zamani et al., 2020; Rao and Daumé III, 2019). However, our problem is even more complex than detecting clarification questions. We are interested in questions aimed at understanding the other’s perspective. Some “clarification questions” might not aim to do that (e.g., “You’re joking, right?” in Table 5).

Our categories have also been relevant in other work: For example, AFC and paraphrasing/summarizing operationalized in the AOP category have been found to be critical to establish a common ground of shared knowledge (Clark and Brennan, 1991; Clark, 1996; Krauss and Fussell, 1991) and increasing understanding of each other (Rogers and Farson, 1957).

A related task is empathy detection (Khanpour et al., 2017; Xiao et al., 2012; Gibson et al., 2015; Zhou et al., 2021; Buechel et al., 2018; Barriere et al., 2022; Lahnala et al., 2022). For example, Gibson et al. (2015) predicted the empathy level of therapists in motivational interviews and defined empathy as: “the extent to which the therapist understands and/or makes an effort to grasp the client’s perspective”. Furthermore, Sharma et al. (2020) developed a framework to measure expressed empathy in online mental health dialogs. Two categories, Explorations (acquiring more information about others’ internal states) and Interpretations (expressing understanding of states of others), relate to perspective-getting but are still too far removed from the setting of online discussions.

3 Operationalizing Text-Based Perspective-Getting

Perspective-getting refers to directly asking about (Eyal et al., 2018; Damen et al., 2021) or listening to (Eyal et al., 2018; Kalla and Broockman, 2023) other person’s perspective (Eyal et al., 2018; Kalla and Broockman, 2023), for example, their experiences, thoughts, feelings, beliefs, or preferences. In lab experiments, perspective-getting has typically been operationalized (1) as whether the perspective-getter asked about the perspective of the other (Eyal et al., 2018; Damen et al., 2021) and (2) as whether the perspective-getter summarized the other’s perspective (Bruneau and Saxe, 2012; Kalla and Broockman, 2023).

Based on (1) and (2), we introduce a codebook (see Appendix A) for the text-based detection of perspective-getting with two distinct categories: Asking for Clarification of the other’s perspective (AFC) or Attending to the Other’s Perspective (AOP), for example, through summarization. Examples of messages displaying each of the two perspective-getting categories can be found in Table 1. While each category could be more or less pronounced in a text, we represent them as a binary value in this work: present or absent. To the best of our knowledge, we are the first to operationalize text-based perspective-getting. Although our categories are intended to be generally applicable, the added examples in the annotation codebook are specific to the Wikipedia discussion pages. We now discuss our categories in more detail.

Attending the Other’s Perspective (AOP) This category is about whether the text message attends the other’s perspective (AOP). The manifestation of AOP can range from quoting or summarizing the original message (Kalla and Broockman, 2023), to otherwise mentioning or referencing specific parts of the message when engaging with the other’s perspective (e.g., by taking a stance on it). Perspective-getting and -taking literature has shown that explicitly attending the other’s perspective (e.g., summarizing or paraphrasing the other) increases the accuracy of understanding the other’s perspective (Bruneau and Saxe, 2012; Kalla and Broockman, 2023). Note that AOP includes less prevalent (e.g., simply referencing and acknowledging the other’s perspective) and more prevalent forms (e.g., actively engaging with the other perspective). In this work, both forms are part of this category; we leave more fine-grained distinctions to future work.

Type		initial	no messages 1000+ words	no errors	only complete discussions	ours
✓	# messages	4,441	4,181	3,693	1,339	263
	# discussions	217	208	203	96	20
✗	# messages	133k	123,712	98,491	14,419	1,411
	# discussions	9,006	8,768	8,711	2,834	263

Table 2: *Data Pre-Processing Steps*. We display the number of discussions and messages that remain for the escalated (✓) and non-escalated discussions (✗), after the following pre-processing steps: (1) removing messages with more than 1000 words, (2) removing duplicates or dangling messages (no errors) and (3) removing incomplete discussions. Out of these, we annotated a selection of messages (ours) that showed disagreement.

Asking for Clarification (AFC) The second category is about actively “getting” or understanding the perspective of the other. This is mostly represented through asking questions (as operationalized in previous perspective-getting literature (Damen et al., 2021; Eyal et al., 2018)). We allow the “questions” to be implicit (e.g., “I think you meant ...” instead of “Did you mean ...?”). Asking about the other’s perspective can lead to better understanding of the other’s perspective (Damen et al., 2021).

4 Dataset

WikiDisputes We use our codebook to annotate perspective-getting in the *WikiDisputes* dataset from De Kock and Vlachos (2021). This dataset is derived from Wikipedia talk pages, where editors discuss possible changes to Wikipedia articles. WikiDisputes consists of 133,019 messages in *non-escalated* discussions and 4,441 messages in *escalated* discussions, i.e. where participants asked for the discussion to be mediated by a moderator. When editors do not reach consensus, the conflict is escalated and resolved by arbitration.¹ Previous research shows that when task-related conflicts escalate, it decreases the group performance related to editing the Wikipedia page (Arazy et al., 2013).

Data filtering We performed several steps to filter the dataset. See Table 2 for the number of remaining messages and discussions after each step.² First, we removed messages with more than 1000 words to reduce annotation time (the mean number of words in a message in the full dataset was 86.19). Second, we removed message duplicates and messages that replied to a non-existent initial message (i.e., the id of the ‘parent’ message points to a message that is not in the dataset). Third, because of

¹https://en.wikipedia.org/wiki/Wikipedia:Dispute_resolution

²The initial #discussions differs from De Kock and Vlachos (2021) because they used different inclusion criteria.

our interest in discussion outcomes we removed incomplete discussions, by removing all discussions in which at least one message was removed in the first two steps. This step led to a large reduction in the number of discussions (Table 2).

Finally, a message was selected for annotation if (i) the message was a reply to a discussion partner’s message (so we did not evaluate topic-starters’ initial messages) and (ii) if the message demonstrated a form of disagreement. When there is agreement between the messages, we assume no perspective-getting to be necessary as the ‘perspective’ is already shared.³ If a person wrote more than one message in a row and directed them to the same user, these messages were merged together.

Annotating perspective-getting Based on our codebook, two researchers independently annotated the WikiDisputes messages. Both annotators were male, one being a master’s student in computational social science and the other an assistant professor in social science. The annotators each independently labeled 165 messages across three iterations. They labeled the perspective-getting of a message with respect to the previous replied-to message. After each iteration the messages that were rated differently were discussed. While the conceptual definitions of the categories remained the same, we used the insights from these discussions to adjust the illustrative examples in the codebook. The 165 messages were not used for the subsequent data analysis. In the final iteration, the Cohen’s kappas for the separate categories were: AOP = 0.78 and AFC = 0.71. Our values indicate substantial to excellent annotator agreement (Landis and Koch, 1977).

After the first three iterations, the two annotators labelled the messages separately. In our final

³Perspective-getting can also take place in discussions when participants already agree. However, such cases were not the focus of this study.

	\sum	AOP			AFC		
	F1	P	R	F1	P	R	F1
LR	0.68	0.95	0.73	0.83	0.54	0.51	0.52
RB	0.76	0.78	0.71	0.74	0.78	0.78	0.78

Table 3: *Classification performance.* Performance of Logistic Regression (LR) and RoBERTa (RB) on the two perspective-getting categories. P stands for precision, R for recall and \sum for macro-average. The results for RoBERTa are means from three runs with different seeds. Standard deviation is omitted as it stays consistently ≤ 0.03 . Best F1 performances are boldfaced.

dataset, AOP is present in 85% of the messages (n=1429), while AFC is only present in 25% of the messages (n=421). In total, we annotated 1,411 messages from the non-escalated and 263 from the escalated discussions.

5 Predicting Perspective-Getting

We now use our annotated dataset to train and test classification models to predict the two perspective-getting categories.

5.1 Classifiers

As a message can contain both perspective-getting categories, we train binary classifiers to predict the presence of each category separately. We experiment with a logistic regression model using a bag-of-words representation and a fine-tuned RoBERTa model. For the two perspective-getting categories, we divide the dataset into different training (64%), validation (16%) and test sets (20%). We do this to keep the same proportion of present and absent categories in the train, development and test sets.

We use logistic regression with L2 regularization as a baseline, implemented using scikit-learn (Pedregosa et al., 2011). Since Litvak et al. (2016) found that specific words can signal perspective-taking, we use the raw frequency of unigrams as features. We remove stop words using a list from the NLTK library (Bird and Loper, 2004). The punctuation signs {‘?’, ‘!’, ‘.’} are treated as separate tokens and we replace all numbers with a <num> token. To penalize the model inversely proportional to the class frequency, we set the ‘class weight’ parameter to ‘balanced’. We set the regularization parameter to 0.1, after experimenting with $\{10^{-3}, 10^{-2}, \dots, 10^2\}$ on the development set.

We fine-tune a RoBERTa model (Liu et al., 2019), with the default hyper-parameters: a learning rate of $5e^{-5}$, 500 warmup steps, and a dropout

	AOP		AFC	
mention	0.41	?		1.22
word	0.37	wikipedia		-0.55
adding	-0.36	work		0.42
edit	-0.36	explain		0.38
know	0.35	feel		0.34

Table 4: *Logistic Regression Coefficients.* Features with the highest predictive weight derived from the Logistic Regression models for the Perspective-Getting Categories: Attending the Other’s Perspective (AOP) and Asking for Clarification (AFC).

probability of 0.1. We use the Adam optimizer, with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e^{-8}$ and a weight decay of 0.1. We use five training epochs based on results on the development set. We select the best performing model by evaluating on every 50 steps on the development set. We report the average results after fine-tuning on three random seeds.

5.2 Results

We evaluate the models based on precision, recall and F-1 scores. We report the macro average due to the uneven class distribution in our dataset, see Table 3. On average, RoBERTa outperforms the logistic regression model (see the F1 macro avg). RoBERTa performs worse for AOP (F1 of 0.74) than for AFC (F1 of 0.78). In contrast, logistic regression performs better for AOP (F1 of 0.83).

Feature and error analysis Table 4 shows highly-weighted features of our Logistic Regression model. Some of these features are specific to the Wikipedia-based training dataset, like *wikipedia* (AFC) and *edit* (AOP). The question mark, the most important feature for the AFC category, seems like a general sign of an attempt to clarify the other’s position. However, a question mark alone is not sufficient for predicting a genuine ask for clarification in many cases (18% of non-AFC utterances include a question mark, e.g., “Is this really difficult to understand for you??!?” and 38% of AFC utterances do not include a question mark, e.g., Table 1).

For RoBERTa, we manually analyzed a sample of instances that the model predicted incorrectly (Table 5). In the first example, the statement “If there is no doubt it is a polish dialect” could be understood as the author expressing interest in the other’s perspective or a request to the previous author to explain their position. However, it could

Cat	Example	P	GT
AFC	<p>Message: There is no doubt it is a Polish dialect. One can doubt only by not knowing a history and listening to today's dialect. But go back to the beginning of the 20th century and read some texts written in the dialect from that period and you'll clearly see it is a Polish dialect with very little Czech influence. Today it sounds like a transitional dialect because of the presence (from 1920) of local people in the Czech state, which heavily linguistically influenced the dialect. But let's move some 500 metres to the Polish side of the border and you will hear more original version of the dialect (but more tainted by the correct Polish language). Btw. read pl:Dialekt śląski. -</p> <p>Reply: If there is no doubt it is a Polish dialect, I am sure you will find reliable sources to support the statement that it is a Polish dialect and not anything else. If you do not, the text should be altered. Before you support the statement by references or the article is altered, the NPOV tag should remain.</p>	0	1
AOP	<p>Message: NPOV</p> <p>Yes this is serious. The meanings of Abort Retry and Fail in this message were actually distinct. Abort would cause the program to return a TRUE value, fail would cause it to return a false value, and retry would simply retry the operation to see if something changed (like a disc being placed in a drive). Again, I'm totally serious about this concern.</p> <p>Reply: You're joking, right? If there's a factual inaccuracy in the article then state your reason here and make the change. I see no POV issues at all.</p>	1	0

Table 5: *Incorrect Model Predictions.* Examples of incorrectly classified messages on AFC and AOP categories, where the model predictions (P) differ from the annotation (GT) with high confidence.

also be read as an uptake of the previous position (i.e., AOP) without genuine interest in further explanation from the previous author. Similar to other areas in NLP, our task might be subject to plausible label variations (Pavlick and Kwiatkowski, 2019; Uma et al., 2021; Jiang and de Marneffe, 2022). However, overall, we reached reasonable agreement between annotators – note though that agreement was lower for AFC than for AOP.

For the second example, the model incorrectly predicted AOP. This prediction is understandable as it sounds like the responding author is paraphrasing the previous author's message by addressing the "factual inaccuracy" mentioned by them. However, when looking at the reply in context (i.e., in relation to the message it is answering to), it becomes clear that the responding author did not directly engage with anything specific the previous author said. When adding even more context, this becomes even clearer: The subsequent reply of the original author starts with "No, I'm sadly not joking. And no, it's not regarding factual inaccuracy [...]". Note that RoBERTa was only fine-tuned on the reply and did not see the original message. This might improve performance in future work.

6 How perspective-getting is associated with discussion course and outcomes

Perspective-getting may influence online discussions. First, conflict management and negotiation scholars found that exchanging information between negotiators positively influenced the negotiation process and outcomes (Trötschel et al.,

2011; Thompson and Hastie, 1990; Weingart et al., 1993; Adair and Brett, 2005). Second, political scientists demonstrated that exchanging narratives between in and out-group members in interpersonal conversation may reduce negative attitudes between the groups (Kalla and Broockman, 2020). Recently, Kalla and Broockman (2023) found that perspective-getting reduced negative stereotypes in inter-group interaction. In sum, there are indications that perspective-getting is an effective strategy to reduce conflict. We therefore now use our annotated dataset to study the influence of perspective-getting on online discussions' outcomes.

6.1 Hypotheses

Perspective-getting and escalations to a moderator In Wikipedia discussions, participants can request a moderator when they can not resolve a discussion by themselves (De Kock and Vlachos, 2021). Morrell (2010) suggest a relationship between interpersonal understanding and having constructive discussions without conflicts between citizens. Further, perspective-getting may reduce conflicts between opposing groups by decreasing thinking in stereotypes (Kalla and Broockman, 2023). Hence, our first hypothesis (H1) is that higher perspective-getting in a discussion are related to a lower rate of escalations to a moderator.

Initial perspective-getting and the course of the discussion Scholars have demonstrated how important the first message in an online conversation is for the course of the conversation (Zhang et al., 2018). Additionally, previous studies have found

that perspective-getting is related to increased positive intentions towards the other and their social group (Kalla and Broockman, 2023; Pagotto, 2010). As perspective-getting is considered as a more effective strategy to understand others (Damen et al., 2021; Eyal et al., 2018; Kalla and Broockman, 2023), we expect a similar influence on the course of the discussion: perspective-getting breeds perspective-getting. Therefore, our second hypothesis (H2) is that subsequent levels of perspective-getting in a discussion are positively related to higher perspective-getting in the first reply. Consequently, the null hypothesis states that the aforementioned relationship is insignificant.

6.2 Control Variables

The course and outcomes of discussions on Wikipedia do not only depend on perspective-getting behavior. There are several factors that need to be controlled for when testing hypotheses. First, discussions may differ in deliberation: how much effort discussants put in providing support for the arguments. Deliberation among discussants is known to decrease the likelihood of escalation. Second, discussions vary in the tone of how discussants communicate with each other. A disrespectful tone, e.g. with insults, is a typical cue for content moderation (Stockinger et al., 2023). Third, discussions may vary in how controversial the topic is. Controversial discussions on Wikipedia, e.g. about religion, are more likely to escalate and require moderation (Yasseri et al., 2014).

Providing Support of own Arguments (PSA)

We control for whether the discussants provide evidence for their own arguments while replying to the original message, which is a sign that they take the effort to justify their claims in a discussion. It is a key measure of deliberation in online discussions (Lundgaard and Etter, 2022), which may decrease inter-group conflict in communication. We annotated PSA along the same process as the perspective-getting categories. See an example in Table 6. Cohen’s kappa was 0.86. In the annotated data, 38% (637) of the messages show PSA compared to 62% (1037) of the messages where PSA was absent.

Respectful Tone of Voice (RTV) We control for the extent to which discussants exercised a neutral to respectful tone of voice towards the other. Tone of voice has been suggested to be an important cue to recognize genuine interpersonal understanding,

as it signals a discussant’s intention to engage with other person’s perspective (Nilsen and Bacso, 2017; Keysar, 1994). See an example in Table 6. The Cohen’s kappa was 0.75. In the annotated data, 79% (1325) of the messages show RTV.

Topic controversiality Studies have shown that how controversial a discussion topic is can influence discussion course and outcomes, such as escalation (Rad and Barbosa, 2012; Alashri et al., 2015). We therefore include how controversial a topic is as a control variable to our analyses. Previous research made a ranking of the most controversial topic categories on Wikipedia (Yasseri et al., 2014). This research identified and ranked controversial topic categories on Wikipedia: politics, countries, religion, history and social issues. We used this ranking to develop an ordinal scale to measure *topic controversiality*. This scale ranged from 5 for a discussion thread about the most controversial category (politics) to 0 for discussion threads about a non-controversial topic category (e.g., music). We developed a codebook for this scale to label the discussion threads in our dataset with the topic categories. For example, the Wikipedia discussion thread “List of sovereign states”⁴ was labeled with the countries category, and received a value of 4 as countries is the second most controversial topic category in the ranking (Yasseri et al., 2014). Two annotators used the codebook to label the topics of all discussion threads into topic categories (N=336). The intraclass correlation of a doubly coded random sample (N=69) was very good (0.77), implying a high level of agreement between annotators.

6.3 Results

Perspective-getting and escalations to a moderator (H1) We first conducted the Levene’s test (Schultz, 1985) to test the t-test’s equality of variance assumption for the perspective-getting of messages in escalated and non-escalated conversations. The test indicated equal variance ($F = 0.5$, $p = 0.82$). We then conducted a one-way ANCOVA to test our first hypothesis. ANCOVA (analysis of covariance) is a statistical method that adjusts for control variables to isolate the influence of categorical predictors (escalation) on an outcome variable (perspective-getting). The null hypothesis stated that the perspective-getting variables, AOP and

⁴https://en.wikipedia.org/wiki/List_of_sovereign_states

Category	Example
Respectful Tone of Voice (RTV)	✓ Message: The images are of poor quality and editorially meaningless. They will continue to be removed until better images or explanations can be found of their editorial utility. Reply: They are the best images out there. Please do not remove the pictures. They are all relevant to the article.
	✗ Message: I think you're lost. Democrat Underground is over here. Reply: [USERNAME], It's disgusting to know that there are people with a bellow 20 IQ like you on the internet. I really hope you're not a U.S. citizen. What a stain on this country you are!
Providing Support of own Arguments (PSA)	✓ Message: One of the problems with your proposal is that it never describes how popular Interstate Football was in Victoria. I am willing to change the line to (But Interstate Football mostly though out history was very popular in Victoria), to describe again that not always, but mostly though out history it was very popular in Victoria. But I won't support not describing what is a fact, and supported in many places that Interstate Football was mostly though out history very popular in Victoria. Reply: Read this article from 1927 and this article from 1944 .
	✗ Message: And how is that relevant to this article or this discussion? This is not a forum. Cheers, Reply: People die from censorship in China. Cheers as you put it.

Table 6: *Control Variables*. We display two example messages manifesting (✓) or not manifesting (✗) the considered control variable.

AFC, have higher scores for escalated messages than for non-escalated messages or there was no difference. Two ANCOVA tests were conducted to investigate the effect of AOP and AFC separately.

For AOP, the ANCOVA results did not reject the null hypothesis. While controlling for topic controversiality, PSA and RTV scores, there was no significant difference between non-escalated ($M = 0.86$, $SD = 0.35$, 95% CI [0.84, 0.87]) and escalated discussions ($M = 0.84$, $SD = 0.36$, 95% CI [0.8, 0.89]): $F(1,1668) = 0.12$, $p = 0.73$, $\eta_p^2 < 0.01$, $N = 1,674$. For AFC, the ANCOVA results did not reject the null hypothesis. There was also no significant difference between non-escalated ($M = 0.26$, $SD = 0.44$, 95% CI [0.24, 0.28]) and escalated discussions ($M = 0.21$, $SD = 0.41$, 95% CI [0.16, 0.26]): $F(1,1668) = 0.64$, $p = 0.42$, $\eta_p^2 < 0.01$, $N = 1,674$.

Our hypothesis is thus not supported: perspective-getting is not higher in non-escalated conversations than in escalated ones. This contrast with previous research on perspective-getting's positive influence on negative inter-group attitudes (Kalla and Broockman, 2023). First, perspective-getting may play less of a role because the goal on Wikipedia is to reach agreement (Damen et al., 2021; De Kock and Vlachos, 2021). Second, De Kock and Vlachos (2021) found that the best predictors to avoid escalation were politeness and collaboration. We controlled for RTV, a related concept to politeness, possibly leading to reduced effects. Last, we did not consider previous discussions between editors that may increase the need for third-party escalation.

Initial perspective-getting and the course of the discussion (H2)

For this hypothesis, only discussions where the first reply message disagreed with the opening post were included, resulting in 262 out of 399 annotated discussions. A multivariate regression analysis was conducted to test the hypothesis. The effect of AOP and AFC were tested separately, while controlling for topic controversiality. The regression analysis showed that the relationship between the AOP score of the first reply and the whole discussion is significant ($b = 0.9$, $t(259) = 20.48$, $p = 0.04$, 95% CI [0.06, 0.17]). Topic controversiality had no significant relationship with the AOP score ($b = -0.08$, $t(259) = -1.13$, $p = 0.26$, 95% CI [-0.24, 0.07]). The effect of AFC was not significant: $b = 0.03$, $t(259) = 1.69$, $p = 0.47$, 95% CI [-0.06, 0.12]). H2 is thus partially supported: the level of AOP in the discussion are positively associated with AOP in the first reply. This association was not found for AFC.

Since perspective-getting could have a diminishing influence over time, we tested the curvilinear association of the summated perspective-getting scores (sum of AOP and AFC) during the course of discussions. A hierarchical multiple regression analysis was conducted with the mean perspective-getting score of a discussion as a dependent variable, and perspective-getting scores of the first reply and topic controversiality as control variables. The hierarchical multiple regression revealed that the perspective-getting score of the first reply contributed significantly to the regression model, $F(1,260) = 9.99$, $p = 0.002$) and accounted

for 3.7% of the variation of the mean perspective-getting score of a discussion. Adding the quadratic polynomial to the model explained additional 0.4% of the variation, however, this change was not significant: $F(2, 259) = 5.61, p = 0.04$. Introducing the cubic polynomial explained 1% of additional variance, $F(3, 258) = 3.74, p = 0.01$, but this change also was not significant. Hence, we found evidence that perspective-getting in a first reply has a diminishing influence on perspective-getting in the discussion over time.

Although we measured perspective-getting instead of perspective-taking, these findings on H2 are partially in line with previous studies on the positive influence of empathetic reactions on harmonious communication (Pagotto, 2010) and the inhibiting influence on interpersonal aggression in discussions (Richardson et al., 1994).

7 Limitations

First, annotation was sometimes challenging due to characteristics of our data. Some replies were intended for a different user, such as the opening poster of the discussion, rather than the user who was directly replied to, which could affect the annotation of AOP. Furthermore, users sometimes added links to the text with the ‘insert link’ function. These links, however, are not visible in the dataset, which could affect the annotation of the PSA control variable, as links could lead to sources that do not support one’s position.

Second, perspective-getting can be operationalized in a more fine-grained way than our binary categories. For example, just acknowledging a perspective (e.g., “I understand that your view is X. But here is what I think ...”) could be seen as less AOP than actually engaging with it (e.g., “It seems the point you are making is X. I understand how this might come from the belief that ...”).

Third, the number of annotated messages is relatively low. Due to the complexity of perspective-getting, we opted for careful annotation by expert annotators instead of item quantity. Due to the small number of escalated discussions in the original dataset, the number of escalated discussions is especially small (c.f. Table 2), possibly influencing the generalizability of the testing of hypothesis 1.

Fourth, the generalizability of our study is limited, as perspective-getting might manifest differently in other domains. For example, Wikipedia editors have to come to an agreement, while so-

cial media users discussing presidential candidates do not. Further, most replies in our dataset will in some form disagree with the previous message, including those with high perspective-getting. Still, our codebook was developed based on theoretical work and should be transferable to other settings, possibly with some minor changes.

Fifth, our data sample is in English and is likely skewed to specific demographics. Our classifiers are therefore likely to be biased towards these same demographics in performance. Furthermore, our classifiers could latch on to spurious features. We do not recommend using the classifier “out of the box” on datasets that are not comparable.

8 Conclusion

In this study, we measured perspective-getting in Wikipedia discussions. Based on theory, we developed a codebook to annotate messages in Wikipedia discussions on two perspective-getting categories, and then fine-tuned a RoBERTa model to predict the categories. In our dataset, perspective-getting was not higher in non-escalated discussions. However, discussions starting with a post attending the other’s perspective are more often followed by responses that also attend to other’s perspective.

With this work, we hope to contribute to studies of how and when strategies for interpersonal understanding can reduce polarization in online discussions (Kwon and Cho, 2017; Shmargad et al., 2022), as perspective-getting allows people to understand, consider and adopt the opposite position in discussions (Eyal et al., 2018; Kalla and Broockman, 2023; Damen et al., 2021). Future research could build on our work to study the relation between perspective-getting and online polarization, while considering the contextual and causal limitations of our study.

Acknowledgements

We thank the anonymous reviewers. This research was supported by the “Digital Society - The Informed Citizen” research programme, which is (partly) financed by the Dutch Research Council (NWO), project 410.19.007.

References

Wendi L Adair and Jeanne M Brett. 2005. The negotiation dance: Time, culture, and behavioral sequences in negotiation. *Organization Science*, 16(1):33–51.

- Saud Alashri, Sultan Alzahrani, Lenka Bustikova, David Siroky, and Hasan Davulcu. 2015. What animates political debates? Analyzing ideological perspectives in online debates between opposing parties. In *Proceedings of the ASE/IEEE International Conference on Social Computing (SocialCom-15)*.
- Ofer Arazy, Lisa Yeo, and Oded Nov. 2013. Stay on the Wikipedia task: When task-related disagreements slip into personal and procedural conflicts. *Journal of the American Society for Information Science and Technology*, 64(8):1634–1648.
- Carolyn M Axtell, Sharon K Parker, David Holman, and Peter Totterdell. 2007. Enhancing customer service: Perspective taking in a call centre. *European Journal of Work and Organizational Psychology*, 16(2):141–168.
- Valentin Barriere, Shabnam Tafreshi, João Sedoc, and Sawsan Alqahtani. 2022. WASSA 2022 shared task: Predicting empathy, emotion and personality in reaction to news stories. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 214–227, Dublin, Ireland. Association for Computational Linguistics.
- Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.
- Laura W Black, Howard T Welser, Dan Cosley, and Jocelyn M DeGroot. 2011. Self-governance through group discussion in Wikipedia: Measuring deliberation in online groups. *Small Group Research*, 42(5):595–634.
- Emile G Bruneau and Rebecca Saxe. 2012. The power of being heard: The benefits of ‘perspective-giving’ in the context of intergroup conflict. *Journal of experimental social psychology*, 48(4):855–866.
- Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4758–4765, Brussels, Belgium. Association for Computational Linguistics.
- Herbert H Clark. 1996. *Using language*. Cambridge university press.
- Herbert H Clark and Susan E Brennan. 1991. Grounding in communication.
- Debby Damen, Monique MH Pollmann, and Teri-Louise Grassow. 2021. The benefits and obstacles to perspective getting. *Frontiers in Communication*, 6:611187.
- Debby Damen, Marije van Amelsvoort, Per van der Wijst, and Emiel Kraemer. 2019. Changing views: the effect of explicit perception-focus instructions on perspective-taking. *Journal of Cognitive Psychology*, 31(3):353–369.
- Debby Damen, Per van der Wijst, Marije van Amelsvoort, and Emiel Kraemer. 2020. Can the curse of knowing be lifted? the influence of explicit perspective-focus instructions on readers’ perspective-taking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(8):1407.
- Christine De Kock and Andreas Vlachos. 2021. I beg to differ: A study of constructive disagreement in online conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2017–2027, Online. Association for Computational Linguistics.
- Christine De Kock and Andreas Vlachos. 2022. How to disagree well: Investigating the dispute tactics used on Wikipedia. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3824–3837, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing (IWP2005)*.
- Wafaa S. El-Kassas, Cherif R. Salama, Ahmed A. Rafea, and Hoda K. Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.
- Nicholas Epley, Boaz Keysar, Leaf Van Boven, and Thomas Gilovich. 2004a. Perspective taking as egocentric anchoring and adjustment. *Journal of personality and social psychology*, 87(3):327.
- Nicholas Epley, Carey K Morewedge, and Boaz Keysar. 2004b. Perspective taking in children and adults: Equivalent egocentrism but differential correction. *Journal of experimental social psychology*, 40(6):760–768.
- Tal Eyal, Mary Steffel, and Nicholas Epley. 2018. Perspective mistaking: Accurately understanding the mind of another requires getting perspective, not taking perspective. *Journal of personality and social psychology*, 114(4):547.
- Adam D. Galinsky, William W. Maddux, Debra Gilin, and Judith B. White. 2008. Why it pays to get inside the head of your opponent: The differential effects of perspective taking and empathy in negotiations. *Psychological Science*, 19(4):378–384. Publisher: SAGE Publications Inc.
- James Gibson, Nikolaos Malandrakis, Francisco Romero, David C. Atkins, and Shrikanth S.

- Narayanan. 2015. Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms. In *Proc. Inter-speech 2015*, pages 1947–1951.
- Yun He, Zhuoer Wang, Yin Zhang, Ruihong Huang, and James Caverlee. 2020. PARADE: A new dataset for paraphrase identification requiring computer science domain knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7572–7582, Online. Association for Computational Linguistics.
- Claire Hughes and Sue Leekam. 2004. What are the links between theory of mind and social relations? review, reflections and new directions for studies of typical and atypical development. *Social development*, 13(4):590–619.
- Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.
- Joshua L Kalla and David E Broockman. 2020. Reducing exclusionary attitudes through interpersonal conversation: Evidence from three field experiments. *American Political Science Review*, 114(2):410–425.
- Joshua L Kalla and David E Broockman. 2023. Which narrative strategies durably reduce prejudice? Evidence from field and survey experiments supporting the efficacy of perspective-getting. *American Journal of Political Science*, 67(1):185–204.
- Aikaterini-Lida Kalouli, Rebecca Kehlbeck, Rita Sevastjanova, Oliver Deussen, Daniel Keim, and Miriam Butt. 2021. Is that really a question? Going beyond factoid questions in NLP. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 132–143, Groningen, The Netherlands (online). Association for Computational Linguistics.
- Jenna Kanerva, Filip Ginter, Li-Hsin Chang, Iiro Rastias, Valtteri Skantsi, Jemina Kilpeläinen, Hanna-Mari Kupari, Aurora Piirto, Jenna Saarni, Maija Sevón, et al. 2023. Towards diverse and contextually anchored paraphrase modeling: A dataset and baselines for Finnish. *Natural Language Engineering*, pages 1–35.
- B. Keysar. 1994. The illusory transparency of intention: Linguistic perspective taking in text. *Cognitive Psychology*, 26(2):165–208.
- Boaz Keysar, Shuhong Lin, and Dale J Barr. 2003. Limits on theory of mind use in adults. *Cognition*, 89(1):25–41.
- Hamed Khanpour, Cornelia Caragea, and Prakhar Biyani. 2017. Identifying empathetic messages in online health communities. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 246–251, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Robert M Krauss and Susan R Fussell. 1991. Perspective-taking in communication: Representations of others’ knowledge in reference. *Social cognition*, 9(1):2–24.
- K Hazel Kwon and Daegon Cho. 2017. Swearing effects on citizen-to-citizen commenting online: A large-scale exploration of political versus nonpolitical online news sites. *Social Science Computer Review*, 35(1):84–102.
- Allison Lahnala, Charles Welch, David Jurgens, and Lucie Flek. 2022. A critical reflection and forward perspective on empathy and natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2139–2158, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.
- Marina Litvak, Jahna Otterbacher, Chee Siang Ang, and David Atkins. 2016. Social and linguistic behavior and its correlation to trait empathy. In *Proceedings of the Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 128–137, Osaka, Japan. The COLING 2016 Organizing Committee.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*.
- Daniel Lundgaard and Michael Etter. 2022. Everyday talk on Twitter: Informal deliberation about (ir-) responsible business conduct in social media arenas. *Business & Society*, 62(6):1201–1247.
- Michael E Morrell. 2010. *Empathy and democracy: Feeling, thinking, and deliberation*. Penn State Press.
- Elizabeth S. Nilsen and Sarah A. Bacso. 2017. Cognitive and behavioural predictors of adolescents’ communicative perspective-taking and social relationships. *Journal of Adolescence*, 56:52–63.
- Lisa Pagotto. 2010. *The interplay of empathy, oneness and perceived similarity in mediating the effects of perspective taking on prosocial responses*. Ph.D. thesis, Università degli Studi di Padova.
- Silvia Pareti, Tim O’Keefe, Ioannis Konstas, James R. Curran, and Irena Koprinska. 2013. Automatically detecting and attributing indirect quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 989–999, Seattle, Washington, USA. Association for Computational Linguistics.

- Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Hoda Sepehri Rad and Denilson Barbosa. 2012. Identifying controversial articles in wikipedia: A comparative study. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration, WikiSym '12*, New York, NY, USA. Association for Computing Machinery.
- Sudha Rao and Hal Daumé III. 2018. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2737–2746, Melbourne, Australia. Association for Computational Linguistics.
- Sudha Rao and Hal Daumé III. 2019. Answer-based adversarial training for generating clarification questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 143–155, Minneapolis, Minnesota. Association for Computational Linguistics.
- Anu Realo, Jüri Allik, Aire Nolvak, Raivo Valk, Tuuli Ruus, Monika Schmidt, and Tiina Eilola. 2003. Mind-reading ability: Beliefs and performance. *Journal of Research in Personality*, 37(5):420–445.
- Deborah R. Richardson, Georgina S. Hammock, Stephen M. Smith, Wendi Gardner, and Manuel Signo. 1994. Empathy as a cognitive inhibitor of interpersonal aggression. *Aggressive Behavior*, 20(4):275–289.
- Carl Rogers and Richard Farson. 1957. *Active listening*. Mockingbird Press LLC.
- Tulika Saha, Aditya Prakash Patra, Sriparna Saha, and Pushpak Bhattacharyya. 2020. A transformer based approach for identification of tweet acts. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.
- Tulika Saha, Apoorva Upadhyaya, Sriparna Saha, and Pushpak Bhattacharyya. 2021. Towards sentiment and emotion aided multi-modal speech act classification in Twitter. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5727–5737, Online. Association for Computational Linguistics.
- Brian B Schultz. 1985. Levene’s test for relative variation. *Systematic Zoology*, 34(4):449–456.
- Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.
- Yotam Shmargad, Kevin Coe, Kate Kenski, and Stephen A Rains. 2022. Social norms and the dynamics of online incivility. *Social Science Computer Review*, 40(3):717–735.
- Andrea Stockinger, Svenja Schäfer, and Sophie Lecheler. 2023. Navigating the gray areas of content moderation: Professional moderators’ perspectives on uncivil user comments and the role of (ai-based) technological tools. *new media & society*, page 14614448231190901.
- Leigh Thompson and Reid Hastie. 1990. Social perception in negotiation. *Organizational behavior and human decision processes*, 47(1):98–123.
- Roman Trötschel, Joachim Hüffmeier, David D Loschelder, Katja Schwartz, and Peter M Gollwitzer. 2011. Perspective taking as a means to overcome motivational barriers in negotiations: When putting oneself into the opponent’s shoes helps to walk toward agreements. *Journal of personality and social psychology*, 101(4):771.
- Don Tugener, Margot Mieskes, Jan Deriu, and Mark Cieliebak. 2021. Are we summarizing the right way? A survey of dialogue summarization data sets. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 107–118, Online and in Dominican Republic. Association for Computational Linguistics.
- Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.
- Kai Wang and Tat-Seng Chua. 2010. Exploiting salient patterns for question detection and question retrieval in community-based question answering. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1155–1163, Beijing, China.
- Laurie R Weingart, Rebecca J Bennett, and Jeanne M Brett. 1993. The impact of consideration of issues and motivational orientation on group negotiation process and outcome. *Journal of Applied Psychology*, 78(3):504.
- Bo Xiao, Dogan Can, Panayiotis G. Georgiou, David Atkins, and Shrikanth S. Narayanan. 2012. Analyzing the language of therapist empathy in motivational interview based psychotherapy. In *Proceedings of The 2012 Asia Pacific Signal and Information*

Processing Association Annual Summit and Conference, volume 2012, pages 1–4.

Jingjing Xu, Yuechen Wang, Duyu Tang, Nan Duan, Pengcheng Yang, Qi Zeng, Ming Zhou, and Xu Sun. 2019. Asking clarification questions in knowledge-based question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1618–1629, Hong Kong, China. Association for Computational Linguistics.

Taha Yasseri, Anselm Spoorri, Mark Graham, and János Kertész. 2014. *The most controversial topics in Wikipedia: A multilingual and geographical analysis*, volume 25. Scarecrow Press.

Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. 2020. Generating clarifying questions for information retrieval. In *Proceedings of The Web Conference 2020, WWW '20*, page 418–428, New York, NY, USA. Association for Computing Machinery.

Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361, Melbourne, Australia. Association for Computational Linguistics.

Ke Zhou, Luca Maria Aiello, Sanja Scepanovic, Daniele Quercia, and Sara Konrath. 2021. The language of situational empathy. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).

A Codebook

HOW TO CODE 1 = the utterance fits with the perspective-getting category (present) 2 = the utterance does not fit with the perspective-getting category (absent). To calculate the perspective-getting score of an utterance: $\#(\text{present})/\#(\text{absent or present})$

DEFINITIONS: Original message: message, which is replied by the current utterance; Current message: the message for which perspective-getting is measured; Author: a user who writes the reply, for which a perspective-getting score is evaluated.

REQUIREMENTS for an utterance pair labeling: 1. The interaction has to be between two separate users. If an interaction is a self-reply, its perspective-getting cannot be assessed and thus the utterance has to be skipped. 2. The interaction must contain a disagreement. If disputers fully agree with each other, the utterance pair should be skipped.

Category & Definition	Present		Absent	
	Description	Example	Description	Example
1. Attending the other's Perspective (AOP). Consideration of the position of an opponent is a crucial component of perspective-taking. For this criterion, it is necessary to ensure that the opponent's position or arguments are attended. An opposite behaviour would be to build own arguments without paying attention to the argumentation of an opponent. The adaptation of the position is not vital for this criterion.	Quoting or rephrasing the original message; Mentioning words/phrases/arguments in the original messages, while trying to argue with them; The answer involves constructive criticism or agreement with the ideas from the original message; Shifting own perspective under the influence of argumentation in the original message; Using phrases such as "I agree with you on...", "You are right about...", etc.	B: ...Technically spoken, Evropariver did right. He cited a source correctly and gave the references. Thus, in my opinion he did nothing wrong there. On the other hand: If someone claims things such as "Western biology begins in Egypt", my alert bells start to ring very loudly. Egypt is not "western"... Second, the author of the book Evropariver used claims that "Five thousand years ago Egyptian priests were already starting to gather a tremendous amount of medical data"...	The arguments or ideas from the original message are ignored. The current message only provides own argumentation. Switching from on-topic to off-topic (e.g., from the page discussion to interpersonal relations)	A: (Crimea is a part of Ukraine) De jure according to Ukrainian law only. No other country's law considers Crimea as Ukrainian land. B: Oh look, another editor with a tiny edit count who is totally not trying to minimize international recognition of Ukraine's territory. (facepalm).
2. Asking for Clarification (AFC). The author is interested in understanding of the alternative position and tries to obtain additional information about the alternative perspective.	Asking clarifying questions; Using phrases such as "If I understood you correctly...", "I think you meant..."	A: Actually, Mein Kampf contained quite a bit of criticism of Judaism: biased, self-serving and wrong-headed, of course, but criticism nonetheless. There's no particular reason that a person cannot be included in both categories which I think is the case here. B: Did he criticize Judaism as a religion or just the Jews as a group? Only in the first case, the category applies.	The position of the opponent is not clarified; Explanations of WikiDisputes rules would not be enough to satisfy this criteria as "present"	A: That's what I wrote according to the source, but it says that real story "Western biology begins in Egypt", please read more carefully when deleting. Also, why did you remove the invention of medicine attributed to Egypt and everything related to it, explain yourself. B: Did you read what I wrote above? Don't pretend you didn't see that. Your source contradicts itself. And why did you remove Lis Magner? Because you don't like it? That's not how we do things around here. And in case you hadn't noticed, this article is about biology, not medicine.
Control variable: Providing Support of own Arguments (PSA). Ensuring that the arguments will be understood by the opponent and acknowledging that the opponent might not know some information.	A person has put efforts to get across own ideas to persuade the opponent; Links to valuable sources are provided to get across own ideas more effectively; Explaining concepts or definitions that could be unknown or sound ambivalent for an opponent; Specific examples or facts instead of abstract ideas are provided; References to the WikiDisputes rules.	A: It's mentioned that psychology is a science in the first sentence, I don't understand the problem. B: I am not saying psychology is not a science (even though that's what I think). No I mean a section detected to the debate on weather psychology is a science or not. I can't access your link, but if you read Science it defines science as "a systematic enterprise that builds and organizes knowledge in the form of testable explanations and predictions about the universe". That describes psychology as well as any other science. Psychology uses the scientific method for much of its research, and has done so for many decades.	Arguments are provided in an abstract manner or not provided at all; People use concepts, phrases or abbreviations that were not mentioned in the original message, without explanations; No attempts to explain or support own position. Promises of evidences in subsequent messages would be considered as "absent" for this criterion Referencing to own experience and nothing else is considered as "absent"	A: can you please explain your concerns about this edit? I was undoing disruptive, autobiographical edits by a sockpuppet IP account. Some of the added content was unduly promotional or supported only by unreliable sources. B: You just deleted everything that someone did saying that it's all shit. I guess, YOU should explain you very aggressive behavior. Do you have any personal issues with the article or the person?
Control variable: Respectful Tone of Voice (RTV). The author uses neutral or respectful language. Exercising offensive or sarcastic language is usually associated with intolerance towards the alternative position.	The utterance has a neutral or respectful tone; No direct signs of hostility of disregard towards the alternative position are shown; Using such phrases as 'please', 'thank you', 'excuse me', etc. in a non-sarcastic manner.	B: I haven't read "Young Stalin" yet. I do not doubt your quote (and there was another source too,anyways) so this isn't any other than asking for a little help (Wikipedia aside), if you please just could tell me in which chapter (number or name) the English pages 310-311 are found, I would be very grateful.	Using hostile, offensive words Using sarcasm Making jokes about the opponent's perspective Reduction of the opponent's position to absurdity.	A: Except I compared the rankings with the whole of Europe, did you read what I said? And for some reason, you want to ignore Africa, Asia and Oceania And please assume good faith. B: ... You don't see anything written in front of you. North America and Latin America have nothing to do with Russia, absolutely nothing. What are you talking about? We'll compare a country to which continent it lies on, not to foreign continents, for example Africa. All this time, you did not contribute a single thing to the discussion, and just kept complaining. I'd suggest you to leave. Not a single thing sounds "sweet" in this sentence.

Table 7: **Annotation rules.** For each developed perspective-getting category, we display the description and example(s) message(s) when the considered perspective-getting category is absent or present. Note: the category names were changed after the annotation procedure for clarity. The initial categories' names were Considering Opponent Perspective (COP) instead of Attending the Other's Perspective (AOP), and Finding Common Ground instead of Asking for Clarification (AFC). As we focus on how interlocutors communicate to acquire more information about the other's perspective, we changed the initial AOP and AFC labels to better embed our research in perspective-getting instead of perspective-taking literature. The instructions of the codebook remained unchanged. Respectful Tone of Voice (RTV) and providing Support of own Arguments (PSA) remained the same. RTV and PSA were initially dimensions of perspective-taking in our study, but were excluded from the perspective-getting dimensions to increase constructive validity. Notations *A* and *B* in the *Example* columns represent original (those that were replied to) and current (the one we label) messages respectively.

Connecting the Dots in News Analysis: Bridging the Cross-Disciplinary Disparities in Media Bias and Framing

Gisela Vallejo¹ Timothy Baldwin^{1,2} Lea Frermann¹

¹The University of Melbourne ²MBZUAI
gvallejo@student.unimelb.edu.au,
{tbaldwin, lfrermann}@unimelb.edu.au

Abstract

The manifestation and effect of bias in news reporting have been central topics in the social sciences for decades, and have received increasing attention in the NLP community recently. While NLP can help to scale up analyses or contribute automatic procedures to investigate the impact of biased news in society, we argue that methodologies that are currently dominant fall short of capturing the complex questions and effects addressed in theoretical media studies. This is problematic because it diminishes the validity and safety of the resulting tools and applications. Here, we review and critically compare task formulations, methods and evaluation schemes in the social sciences and NLP. We discuss open questions and suggest possible directions to close identified gaps between theory and predictive models, and their evaluation. These include model transparency, considering document-external information, and cross-document reasoning.

1 Introduction

The depiction of complex issues in the media strongly impacts public opinion, politics, and policies (Ghanem, 1997; Giles and Shaw, 2009). Because a handful of global corporations own an increasing proportion of news outlets, the reach and impact of biased reporting are amplified (Hamborg, 2020). Although perfect neutrality is neither realistic nor desirable, media bias turns into an issue when it becomes systematic. If the public is unaware of the presence of bias, this can lead to dangerous consequences, including intolerance and ideological segregation (Baly et al., 2020).

Figure 1 illustrates the concepts of framing and media bias adopted in this paper, using the passing of the Respect for Marriage Act as an example. *Framing* refers to the emphasis of selected facts with the goal of eliciting a desired interpretation or reaction in the reader (Entman, 2007). The left-leaning article in Figure 1 leads with an uplifting

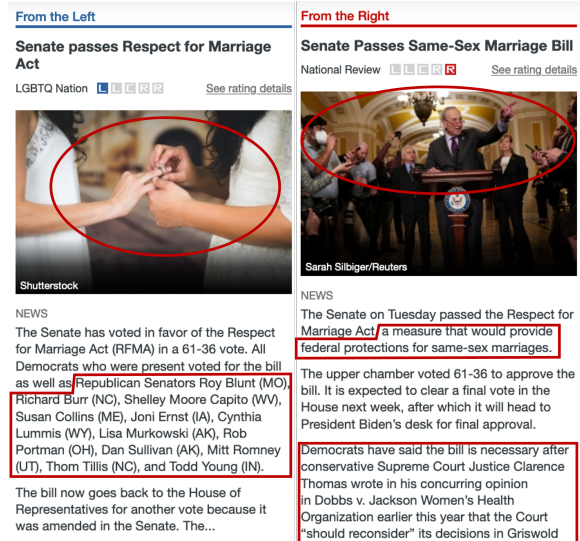


Figure 1: Two articles about the same event written from different political ideologies (Source: allsides.com).

picture of a wedding and emphasizes bill support, evoking a positive framing of new opportunities for same-sex couples; while the right-leaning article focuses on disputes in both image and text, framing the issue in a more negative light. *Political bias* refers to partisan slanted news stories, or the “tendency to deviate from an accurate, neutral, balanced, and impartial representation of ‘reality’ of events and social world” (McQuail and Deuze, 2020), which can be a result of a selected framing. In Figure 1, each document was flagged as far-left and far-right ideological leaning, respectively, on the basis of their publishing media outlets. Political bias is typically deliberate (Williams, 1975) while framing may be inadvertent and caused by external pressures such as space limitations.

Framing and media bias have been under active research in different subfields of the social sciences. Angles of study include the manifestation of frames in the mass media and their effects on public opinion (communication sciences); the impact of frames in groups’ and individuals’ sense-

making of the world (social psychology; sociology) or on their observable behaviour (economics and political science). We focus on the first notion: systematic analyses of framing bias in the mass media, through manual coding, or with NLP technology. In this paper, we will collectively refer to the studies of communication and mass media as social sciences.

With the increasing pace and almost complete digitisation of news reporting there is a need and opportunity to scale the analysis of media bias (Parasie, 2022). Besides, evidence suggests that exposing media bias promotes healthy public debate, aids journalists to increase thoroughness and objectivity, and promotes critical news consumption (Dallmann et al., 2015). We discuss the specific role of NLP in this context in Section 5.

1.1 Contribution and Approach

We relate the NLP research landscape on framing and media bias prediction to typical research questions and hypotheses in the social sciences. We tease out disconnects across disciplines, and make concrete suggestions on how social science approaches can improve NLP methodology, and how NLP methods can more effectively aid social science scholars in their analyses and underpin technology to raise awareness of media bias.

Hamborg et al. (2019) present an overview of traditional and computational approaches to media bias, including detailed definitions of bias types and their emergence in the context of news production. We complement this survey by contextualising recent approaches in NLP with dominant questions and approaches in the social sciences. Ali and Hassan (2022) review computational approaches to modelling framing providing a systematic overview of NLP and machine learning methods. In contrast, we critically review the methodological decisions along the higher-level NLP pipeline: data (Section 4.1), problem formulation (Section 4.2), and evaluation (Section 4.3), link them back to social science methodology, and pinpoint gaps between the two disciplines. We motivate our focus with a case study in Section 3.

We obtained an up-to-date inventory of NLP approaches to media bias and framing, as well as a representative body of corresponding work in the relevant social science disciplines as follows. We collected relevant NLP benchmark data sets (Table 1) and the papers that addressed them for a broad-coverage overview of approaches in the field.

We complement this with social science papers departing from citations in Hamborg et al. (2019). Here, we do not attempt a systematic literature review, but rather present a *representative* body of work across the fields.¹ We excluded papers that a) duplicated methodologies, b) provided redundant definitions, or c) focused on unrelated topics. From this selection process, our final corpus comprises 63 papers (36 framing, 27 media bias), which were considered for further analysis, also listed in Appendix A.

2 Background: Framing and Media Bias

Framing and *politically biased news reporting* are two strategies to systematically promote specific perspectives on contested issues. We note that not every presence of framing is political bias and not all political bias is represented as framing but their intersection can reinforce each other’s impact. They are overlapping concepts which have been addressed jointly or with similar methods in NLP. As such, we include both strategies in this survey.

Framing has been conceptualised variously in different social science disciplines. Prevalent notions of framing include *equivalence framing* – presenting the same logical information in different forms (Cacciatore et al., 2016) – and *emphasis framing* – highlighting particular aspects of an issue to promote a particular interpretation (Entman, 2007). Additionally, framing has been conceptualised as a process (de Vreese, 2005; Entman, 2007; Chong and Druckman, 2007), a communication tool (Scheufele, 1999), or a political strategy (Roy and Goldwasser, 2020). Frames have been conceptualised within different dichotomies. de Vreese (2005) distinguishes *issue-specific* and *issue-generic* frames which apply to only a single or across several issues, respectively. Scheufele (1999) differentiates between *media frames*, as embedded in the political discourse, and *audience frames*, as the reader’s interpretation of an issue. Finally, Iyengar (1991) defines *episodic framing* as portraying an issue with an individual example compared to *thematic framing*, which takes broader context into account. Here, we cover both issue-specific and issue-generic frames and attach to Entman (2007)’s notion of emphasis framing.

While framing is a priori detached from partisan

¹We intentionally depart from the traditional approach of selecting the top N results from a research anthology for a few simple queries, as this would not capture the diversity of works both in terminology and publication venues.

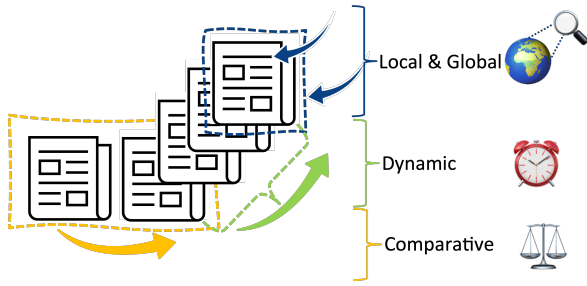


Figure 2: Illustration of the three disconnects: framing is both local and global (blue), dynamic (green) and best identified through comparative analysis (yellow).

views, *political bias* refers to an explicit association of an article or media outlet with a specific political leaning. Both concepts result in biased news reporting, and correspondingly NLP researchers have attempted to address them jointly, either by investigating political framing (Roy and Goldwasser, 2020) or by identifying correlations between framing and partisan slanted articles (Ziems and Yang, 2021). NLP studies have attempted automatic media bias identification under several names, including: hyper-partisan news detection (Kiesel et al., 2019), media bias detection (Spinde et al., 2021b; Lei et al., 2022), identification of biased terms (Spinde et al., 2021a), and political ideology detection (Iyyer et al., 2014; Kulkarni et al., 2018). Their common goal is to detect and classify the bias of a data sample towards a particular political ideology. Many of these approaches naturally relate to investigating *how the story is told* (i.e., framing).

3 Three Disconnects

To illustrate the disconnects between the social sciences and NLP, we use a representative study of media bias from the communication sciences (Hernández, 2018) which investigates the framing of domestic violence in the South China Morning Post. The author formulates two research questions:

1. Framing functions: Are femicides recognized as a problem of domestic violence? What are their causes, and the solutions proposed?
2. Frame narratives: What are the main narratives? And which sources are cited in support?

The first research question considers the *local* written aspects within each news article. Specifically, it studies the causes and solutions presented, grounded in Entman (1993)’s conceptualisation of framing in terms of a problem, its cause, and its solution. The second research question relates these

local aspects to a *global* (cross-document) view by contrasting narratives that present domestic violence as isolated incidents with those that treat it as a societal problem. It further connects the articles to *extrinsic* variables, including the sources used and cultural contexts of the story (e.g. whether the article refers the role of women in the Chinese family or understands domestic violence through the lens of the Confucian philosophy). Furthermore, the study considers articles over an extended period, capturing the *temporal development* of framing and bias. In contrast, current NLP approaches to frame prediction have predominantly adopted a single-label prediction approach per unit of analysis (Baumer et al., 2015; Naderi and Hirst, 2017; Liu et al., 2019), rather than treating frames as structures which could decompose into aspects like cause vs. solution (but see Akyürek et al. (2020); Mendelsohn et al. (2021); Frermann et al. (2023) for recent exceptions). Current approaches furthermore treat units of analysis (sentences, articles) as independent without considering links across documents, across time, or to document-external context. The multi-level and dynamic understanding of bias and framing is fundamental in the social science studies. In sum, we identify three fundamental properties of bias and framing that underpin social science research on bias and framing, and we also visually represent them in Figure 2:

Framing/bias is local and global It is local, because an article can contain several frames, and it is global because understanding the framing of an article may require to aggregate local frames and link them with information such as cited (or omitted) sources, or the outlets’ political leaning.

Framing/bias is dynamic Frames change across time, outlets, countries, and communities. Understanding the *dynamics* of framing can shed light on trends and the impact of a sustained exposure to biased reporting on readers’ opinions.

Framing/bias as a comparative task Media bias and framing are most apparent when directly contrasting articles from different perspectives, places or times (cf., Figure 1). Formulating our task in a comparative way – rather than predicting instance labels in isolation – may improve the quality, reliability and interpretability of predictions.

Only 14.3% of our surveyed papers (N=9) address the global vs local aspect, 9.5% (N=6) explore the dynamics, and 1.6% (N=1) tackle fram-

ing bias as a comparative task over two or more data samples on the same event. The full list of papers and their categorisation can be found in Appendix A. The remainder of this article links these fundamental disconnects to the more practical research design decisions that arise across both disciplines: data, methods and evaluation.

4 A Critical Review of Current Practices in NLP and Social Science

To increase its potential for impact, NLP research needs to reconsider framing and political bias across the entire research pipeline. This includes refining benchmarks, methodologies, and evaluation strategies. In this section, we make recommendations for each component: exploring new aspects of existing datasets, moving beyond single-label classification and incorporating linguistic features as well as external information, and providing transparent and reliable evaluation outputs with error analyses. We critically compare approaches across NLP and the social sciences, pointing out discrepancies together with practical suggestions for future work.

4.1 Datasets

Social science studies are characterised by carefully collated data sets which are, however, typically small in size ($\ll 100$ articles) and manual labels are rarely released to the public. Hence we focus on limitations and opportunities of NLP framing and bias benchmarks in this section. Table 1, lists relevant datasets, along with details on their labels, size, tasks and unit of analysis.

Media bias detection At the *sentence level*, [Lim et al. \(2020\)](#) used crowdsourcing to annotate sentences on 46 English-language news articles about 4 different events with four levels of bias (not-biased, slightly biased, biased, or very biased). [Spinde et al. \(2021b\)](#) released BABE (“Bias Annotations By Experts”), a collection of sentences labelled by experts according to binary categories: biased and non-biased, at the sentence and word levels. [Fan et al. \(2019\)](#) provided the BASIL (“Bias Annotation Spans on the Informational Level”) dataset containing sentence (span) and word-level annotations of political leaning and sentiment (stance) towards entities in the article.

At the *document level*, the Bitterlemons corpus ([Lin et al., 2006](#)), comprises weekly issues about the Palestine–Israel conflict. Each issue contains

articles from Palestinian and Israeli perspectives written by the portal’s editors and guest authors. Despite being intended for document classification, this dataset can be employed to explore framing and political bias, given the documents’ nature of strong bias towards one side of the conflict. Additionally, the web portal AllSides² categorises news outlets into three political ideologies: right, centre, and left (they also offer a finer-grained five-point scale annotation: left, lean left, centre, lean right, right) with the aim to provide all political perspectives on a given story (cf., Figure 1) including expert manual assigned categories at the article level. Several research groups have contributed datasets scraped from AllSides ([Chen et al., 2018](#); [Baly et al., 2020](#); [Liu et al., 2022b](#); [Lee et al., 2022](#)).

Framing At the *headline level*, [Liu et al. \(2019\)](#) released the Gun Violence Frame Corpus (GVFC). It includes headlines about gun violence in news articles from 2016 and 2018 in the U.S., labelled with frames like politics, economics, and mental health. [Tourni et al. \(2021\)](#) released a multi-modal version of the GVFC collection, including the main image associated with each article, and annotations about relevance and framing at the image level.

At the *document level*, the Media Frames Corpus (MFC, [Card et al., 2015](#)) is the currently most extensive frame-labeled data set available. It includes articles from 13 U.S. newspapers on three policy issues: immigration, same-sex marriage, and smoking. This dataset is intended to enable the analysis of policy issue framing, providing annotations at document and span levels with frames like morality, economic, and cultural. More recently, [Piskorski et al. \(2023a\)](#) released a multilingual multifaceted data collection that includes framing as one of the facet with 14 generic framing dimensions at the document level, inspired in the MFC’s annotation. [Ziems and Yang \(2021\)](#) contribute a police violence news articles collection (PVFC) that can be categorised in both domains, media bias and framing. They provide annotations for political leaning: conservative, liberal or none and also entity-centric frames, including the victim’s age, race, and gender.

Opportunities for Future Work. In Section 3, we propose three main aspects to investigate framing and media bias. (1) *Conducting studies at a local and global level.* [McLeod et al. \(2022\)](#) suggest that framing can occur at different textual units in a

²<https://www.allsides.com/about>

Dataset	Categories	Size	Unit of Analysis	Task
Bitterlemons (Lin et al., 2006)	Perspective (Israel, Palestine)	594	Documents	Classification
Flipper (Chen et al., 2018)	Left, Centre, Right	6,447	Documents	Classification
BASIL (Fan et al., 2019)	Liberal, Conservative, Centre; Pos, Neu, Neg	1.2k / 448 300	Spans/Words Documents	Classification
AllSides (Baly et al., 2020)	Left, Centre, Right	34k	Documents	Classification
BiasedSents (Lim et al., 2020)	not-, slightly-, very-, biased	966	Sentences	Classification
BABE (Spinde et al., 2021b)	Biased, Non-biased	3.7k	Sentences	Classification
BIGNEWSALIGN (Liu et al., 2022b)	Left, Centre, Right	1M	Documents	Classification
NeuS (Lee et al., 2022)	Left, Centre, Right	10.6k	Documents	Cross-Doc Summarisation
MFC (Card et al., 2015)	15 Frames	61.5k/ 11.9k	Sentences/ Documents	Classification
GVFC (Liu et al., 2019)	9 Frames	2.99k	Headlines	Classification
Multimodal GVFC (Tourni et al., 2021)	9 Frames	1.3k	Headlines + Images	Classification
PVFC (Ziems and Yang, 2021)	Entity frames; Conservative, Liberal, none	82k	Documents	Entity frame prediction
Narrative Frames (Frermann et al., 2023)	3 entity roles; 5 frames	428	Documents	Multi-label frame prediction
SemEval-2023 Task 3 (Piskorski et al., 2023a)	14 Generic frames	~1k	Documents	Multi-label/-class classification

Table 1: Prominent benchmarks for political bias (top) and framing (bottom). We report size (number of data points), unit of analysis, supported task(s) and labels. All these data sets are in English and most of them U.S. centric.

document. Building on this idea, we propose a shift from single label classification on NLP datasets like AllSides, and Bitterlemons. As a concrete example, these corpora could be used to identify predictive sentences or spans for particular frames of political biases, and investigate commonalities. This can directly inform social scientists in their analyses as well as tools to expose biases to news consumers. Roy and Goldwasser (2020) used point-wise mutual information (Church and Hanks, 1990) over bigrams and trigrams to identify spans but found poor generalisation of the approach. Khanehzar et al. (2021) modelled latent frames at the event level, with not explicit validation. Other specific examples with existing data include: exploring the MFC sentence-level annotations to investigate local framing, and then aggregating these labels to gain a global perspective – an approach that, to our knowledge, has not been done before. Regarding datasets providing sentence-level (BABE) and headline (GVFC) annotation, this can be considered as a local dimension. However, they generalise from the headline to the entire document, which ignores the subtle signals in the local dimension. (2) *The dynamics of framing* on various levels are captured by current data sets: the MFC, BASIL, GVFC and BABE provide article timestamps, supporting diachronic modeling of bias and framing. While some studies exist in this domain (Kwak

et al., 2020; Card et al., 2022), the majority of NLP framing considers articles in isolation. Other dynamics, e.g., across countries, communities or media types (e.g., news vs. blogs) are of central interest in communication studies but less achievable with existing data sets. Constructing cross-language and/or cross-cultural data sets with articles aligned on the event level is an important first step. (3) *Framing as a comparative task*. We propose that researchers explore cross-document differences in their presentation of a specific issue. More concrete, several of the datasets obtained from AllSides include event-level alignment and hence enable comparison across documents on the left–centre–right spectrum at a finer granularity.

4.2 Methodologies

In NLP, researchers have approached media bias as political ideology detection or framing categorisation using different task formulations. The first and most common strategy is *single-label classification*, i.e. assigning a single label to each data point. At the *word level*, Recasens et al. (2013) learn linguistic features from word removal edits in Wikipedia. Spinde et al. (2021a) compared the Euclidean distance of word embeddings to identify biased words in articles from Huffington Post (left wing) and Breitbart News (right wing). And Liu et al. (2021) experimented with identifying and

replacing bias-inducing words with neutral ones using salience scores over word embeddings.

At the *sentence level*, Iyyer et al. (2014) used RNNs to identify political ideology in sentences in congressional debate transcripts and articles from the Ideological Book corpus. Using the BASIL corpus, Hartmann et al. (2019) correlated sentence and document distributions using a Gaussian mixture model (Reynolds, 2009) to identify biased sentences; Chen et al. (2020a) classified biased spans by calculating their probability distributions on news articles; and Guo and Zhu (2022) applied contrastive learning and created sentence graphs to categorise biased sentences. Other researchers translated keywords from GVFC into several languages, and fine-tuned mBERT to classify frames in news headlines in languages other than English (Akyürek et al., 2020; Aksenov et al., 2021).

At the *document level*, there has been substantial work building on the MFC corpus. The task has been approached with RNNs (Naderi and Hirst, 2017), attention and discourse information (Ji and Smith, 2017), and pre-trained transformer models (Khanehazar et al., 2019). Baly et al. (2020) combined adversarial adaptation and adapted triple loss with features like Twitter and Wikipedia information about the readers and the outlet to classify the political ideology of news articles. More recently, Chen et al. (2020b) analysed patterns at different granularities (from word to discourse) to identify media bias and Hong et al. (2023) developed a multi-head hierarchical attention model to identify biased sentences focusing on their semantic and aggregating those for political bias document classification. Scholars have performed similar tasks on languages other than English, e.g. by translating English keywords in MFC to Russian to investigate the U.S. framing in Russian media over 13 years (Field et al., 2018).

Some work has formalized framing/bias detection as *multi-label classification*, typically adopting unsupervised methods like clustering (Ajjour et al., 2019) or topic modelling (Tsur et al., 2015; Menini et al., 2017) which allows to ‘softly’ assign documents to more than one cluster. In a supervised manner, Mendelsohn et al. (2021) employ RoBERTa to classify multiple framing typologies on immigration-related tweets. Similarly, Akyürek et al. (2020) address multi-label framing over headlines using different configurations of BERT. Both works focus on short documents (headlines or articles capped at 280 characters). The very recent

work of Frermann et al. (2023) is the first to address document-level multi-label frame classification. Rather than unstructured, ‘topic-like’ frame detection, some works anchored framing in the depiction of important stakeholders, also referred to as *entity framing* (Ziems and Yang, 2021; Khanehazar et al., 2023).

While we focus on frame and bias *detection*, NLP has also proposed methods for *mitigation*, e.g., by flipping of bias of headlines (Chen et al., 2018) or generating neutral summaries from a collection of biased articles on the same topic (Lee et al., 2022). These applications come with their own sets of methodological and evaluation challenges, as well as ethical risks, and are beyond the scope of this paper. We advocate for the alternative approach of highlighting frames in multiple articles and presenting them side-by-side as illustrated in Figure 1, as a safer and potentially more effective approach in raising awareness of bias and framing.

In the social sciences, approaches tend to be manual, with fewer data samples. One common approach is to *reason across many documents from a high-level perspective*. For example, Chyi and McCombs (2004) design and evaluate a two-dimensional framework (spatial and temporal) to investigate framing changes over time in 170 news articles in American English about a U.S. school shooting event. They manually annotated articles with the signals indicating both of the frame typologies, quantified those annotations and draw conclusions about the temporal and spatial framing behaviour in the inspected articles. Muschert and Carr (2006) assessed the previously-proposed framework based on 290 news documents, and confirmed that the present temporal dimension frame still holds when using data from more than one school shooting. Hernández (2018) analysed the framing of 124 news stories from the South China Morning Post (SCMP) about femicides by manually coding the articles and quantifying those observations. The author explored whether those cases were portrayed as isolated cases or part of a systematic social problem, by manually analysing signals like narratives, sources, and the role of the entities.

Communication science studies often *correlate features of news reports with extra-textual information to formulate or validate their hypotheses*. For example, McCarthy et al. (2008) assess media bias in reporting on demonstrations. They examine media coverage of protests during Belarus’s transition from communism, considering factors like

protest size, sponsors' status, arrests, and their correlation with media coverage. Similarly, [Gentzkow and Shapiro \(2010\)](#) investigate media bias by calculating think tank citation frequencies in media outlets and correlating them with U.S. Congress members mentioning the same groups.

Opportunities for Future Work. There is a stark disconnect between largely *local* approaches to frame modelling in NLP and the focus on *dynamic* and *global* questions explored in framing/bias studies in the social sciences. These arguably more complex questions emerging from the social sciences can guide the development of NLP methodologies. Specifically, capturing subtle signals, including the metaphoric or technical (legal) language use, the correlation with external features, e.g. a report's sources, and the broader cultural context in which an article emerged can enrich news framing and bias analysis. Examples at a linguistic level include enriching framing models with notions of metaphoric ([Chakrabarty et al., 2022](#); [Liu et al., 2022a](#)) or subjective ([Barrón-Cedeño et al., 2023](#)) language. On the cross-document and dynamic level, we propose to address bias and frame classification as a comparative task rather than classifying documents in isolation. This can help *induce* frames from data by analysing axes of largest variation; and can naturally support tools and applications to raise readers' bias awareness by exposing them to contrasting perspectives on the same issue. Contextualising framing models with extra-textual, cultural context is arguably the most challenging gap to fill. While it is tempting to suggest the use of large language models to draw some of these connections, we strongly argue for using them at most as an aid for human domain experts, and to scrutinise any automatic predictions due to the known intrinsic biases in these models.

4.3 Evaluation

We consider two levels of validation: validating data annotations, and validating model predictions.

Validating annotations Validating the quality of labelled data applies to both the social sciences and NLP. In a typical social science study, the distribution of manual labels is the main factor for accepting or rejecting hypotheses. As such, measures for data quality such as inter-coder reliability (ICR) are routinely reported and a core requisite of the study to ensure that the codebook was correctly conceptualised. Coding often includes discussions and

several iterations on trial data ([Hernández, 2018](#)), leading to relatively high ICR scores from carefully trained annotators, often with domain knowledge. For robust NLP model training and validation, reliable annotations are essential. While the assessment of bias or framing are subjective to some extent – as the assessment of framing depends on the annotator's predispositions – the development of *scalable* annotation frameworks that minimise subjectivity is an important open problem.

Validating (model) predictions Social science studies are largely analytical examining labelled data, qualitatively based on manual analysis, and quantitatively based on statistical tests. In contrast, NLP framing studies primarily rely on empirical methods, evaluating through numerical comparisons with ground truth labels. We propose a shift towards deeper insights, assessing a model's ability to capture framing and political bias on a higher, more abstract level, while also fostering fresh insights into the data. Current approaches fall short of drawing inferences from explicit information, such as assessing story objectivity and factuality. These nuanced, graded strategies require more comprehensive metrics than binary accuracy.

Opportunities for Future Work. We particularly suggest the consequent adoption of three levels of evaluation: (1) model performance, (2) error analysis, and (3) measuring model certainty. While the three levels are by no means new, NLP work continues to focus on (1), with (2) and (3) given less thought and rigour. NLP research on media bias would benefit from established standards that guide the error analysis as well as measures of model reliability and (un)certainty. Such standards might include reporting of 'most challenging' classes and/or instances; categorization of errors; as well as exploring reasons for such shortcomings ([Vilar et al., 2006](#); [Kummerfeld and Klein, 2013](#)). Finally, with the increasing impact of NLP technology on the broader public, users of resulting models (be it news consumers or social science researchers), must have access to model confidence scores to assess the reliability of model predictions, as per point (3).

5 Discussion

Harmonising depth and scale The differences in data sets and evaluation between the disciplines naturally follow from their respective goals.

Framing studies in the social sciences aim to uncover the principles underlying framing and its effects through careful, manual analysis of limited amounts of data, typically grounded in theoretical constructs. The primary goal of NLP in the space of media analysis is automation and scalability. Complex annotation of large training data sets as required for supervised approaches is infeasible. Besides, the required structured annotation paradigms would result in sparse observations of label co-occurrence which in turn would require even larger labelled corpora – and exploding annotation costs. Harmonising the goal of scalability with depth and theoretical rigour is a difficult problem (that is not specific to the domain of framing and media bias). One approach towards addressing this problem is the use of semi- or unsupervised approaches, which limit the annotations to evaluation sets of more manageable size. Incorporating small amounts of labelled data with powerful pre-trained models is an obvious methodological approach, however, ensuring the validity of predictions and interplay of biases encoded in these models with the target task at hand is an open and important research problem – particularly in a sensitive domain like media bias analysis.

Feasible yet valid annotation How can we obtain ecologically valid annotations in an efficient way and sufficient quantity? We suggest to follow a common strategy in the social sciences: break articles into self-contained segments, on the event or argument level (Muschert and Carr, 2006). While recent work on argumentation in online debates has followed a similar approach of segmenting contributions into arguments and annotating frames on the argument level (Ajjour et al., 2019), it has not been applied in the news media context. Localised rather than article-level annotations have three advantages: (1) a cognitively easier task for annotators; (2) interpretability through the possibility to provide local, extractive evidence for frame predictions; and (3) a richer document-model of framing that goes beyond the single most likely frame.

Cross-disciplinary expertise for document-external grounding Section 3 pointed to a need for multi-level bias analysis, incorporating local, cross-document and broader cultural contexts. Most NLP work models individual articles without integrating external information or other articles in the collection. A few exceptions exist, includ-

ing Baly et al. (2020) who incorporate readership demographics from Twitter and publisher information from Wikipedia; and Kulkarni et al. (2018) who incorporate article link structure into their models. Both works still model data points in isolation, and fall short of incorporating the more subtle cultural, political or societal contexts that inevitably interact with news framing. We argue for a strong role of cross-disciplinarity and human oversight when incorporating those factors, involving domain experts at every step from formulating research questions to model design, transparency, robustness, and evaluation. Cross-disciplinary projects would guide NLP researchers to develop novel methods that are valid and useful for studying the fundamentals of framing and media bias, and equip social scientists with enlarged data sets of high quality and relevance to enrich their research.

Open data NLP has a strong culture of sharing code and annotated data sets to encourage collaboration and reproducibility. This is less common in the humanities. Sharing this data more explicitly through cross-disciplinary dialogue could provide critical assessment and feedback from domain experts. It could drive research into combining large (and potentially noisier) data with small-scale (but high-quality) data sets from the social sciences, to address increasingly complex questions on the emergence and effects of media biases and framing.

The role of NLP in media bias analysis Despite a surge in data sets and models for automatic analysis of frames and media bias, the *ultimate goal* of these works receives surprisingly little attention. With the broader adoption of NLP methods diverse applications emerge – from supporting social scientists in scaling their research to larger data samples, to tools that highlight (or even edit) biased news to general public news consumers to expose slanted reporting. An explicit notion of goals and applications (and corresponding statement in research papers) will inform model evaluation, risks and ethical concerns to be discussed in the paper. A mandatory adoption of model cards (Mitchell et al., 2019) is one step in this direction. Irrespective of the final application of NLP research, we argue that NLP can contribute safe and valuable tools and methods only if it recognises the complexity of bias and framing both in its data sets and annotations as well as in its evaluation procedures.

6 Conclusion

We surveyed recent work in NLP on framing and media bias, and identified disconnects and synergies in datasets, methodologies, and validation techniques to research practices in the social sciences. Despite the opportunities for NLP to support and scale social science scholarship on media bias, a current oversimplification in conceptualisation, modelling, and evaluation limits the validity and reliability of contributions. We have teased out three disconnects and proposed directions for future work, including: (1) analysing news articles from a local and global perspective, incorporating external non-textual features; (2) taking into account the dynamics of framing and bias across documents, cultures or over time; and (3) tackling the issue of media bias as a comparative task, defining frames on the basis of systematic differences between articles whose origins differ on pre-defined characteristics. This would allow for a more complex characterisation of bias than the currently dominant approach of single-label classification.

Limitations

This survey focuses on media bias and ‘frame building’, i.e. the manifestation of biases and frames in news articles. This constrains the scope of our analysis to mainstream print news outlets; and leaves aside the dimension of ‘frame setting’, i.e. the effects of those frames on the news consumers. Additionally, we are aware that regardless of the approach taken for sampling the body of previous work included in this paper, given the vast literature in the social sciences, there will be remaining bias in our selection. With the aim of mitigating this bias, we point the reader to complementary surveys in this field, e.g. [Hamborg et al. \(2019\)](#) and [Ali and Hassan \(2022\)](#).

Ethics Statement

Identifying framing and political bias in news articles is a sensitive application area, and inevitably influenced by social and structural biases in the academic investigators and the pool of annotators. Datasets and technologies intending to tackle these phenomena comprise the social bias of annotators and researchers developing them in an environment lacking diversity. Besides there is a potential for dual use of models and benchmarks to promote polarisation and misinformation through framing,

rather than reduce it. We see this paper as an opportunity to identify new directions to diversify NLP methodologies and data sets, grounded in best-practices from the media sciences which have been developed for decades. We anticipate that these steps will, together with a better documentation of models and intended use cases, will help to address the above concerns.

Acknowledgements

We thank the anonymous reviewers for their feedback, which significantly improved this paper. We also thank Max Glockner and Vishakh Padmakumar for their constructive suggestions and feedback on this work. This article was written with the support from the graduate research scholarship from the Melbourne School of Engineering, University of Melbourne provided to GV. LF is supported by the Australian Research Council Discovery Early Career Research Award (Grant No. DE230100761).

References

- Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019. [Modeling frames in argumentation](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2922–2932, Hong Kong, China. Association for Computational Linguistics.
- Dmitrii Aksenov, Peter Bourgonje, Karolina Zaczynska, Malte Ostendorff, Julian Moreno-Schneider, and Georg Rehm. 2021. [Fine-grained classification of political bias in German news: A data set and initial experiments](#). In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 121–131, Online. Association for Computational Linguistics.
- Afra Feyza Akyürek, Lei Guo, Randa Elanwar, Prakash Ishwar, Margrit Betke, and Derry Tanti Wijaya. 2020. [Multi-label and multilingual news framing analysis](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8614–8624, Online. Association for Computational Linguistics.
- Mohammad Ali and Naeemul Hassan. 2022. [A survey of computational framing analysis approaches](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9335–9348, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. [We can detect your bias:](#)

- Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991, Online. Association for Computational Linguistics.
- Alberto Barrón-Cedeño, Firoj Alam, Tommaso Caselli, Giovanni Da San Martino, Tamer Elsayed, Andrea Galassi, Fatima Haouari, Federico Ruggeri, Julia Maria Struß, Rabindra Nath Nandi, et al. 2023. The clef-2023 checkthat! lab: Checkworthiness, subjectivity, political bias, factuality, and authority. In *European Conference on Information Retrieval*, pages 506–517. Springer.
- Eric Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. 2015. [Testing and comparing computational approaches for identifying the language of framing in political news](#). In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1472–1482, Denver, Colorado. Association for Computational Linguistics.
- Michael A Cacciatore, Dietram A Scheufele, and Shanto Iyengar. 2016. The end of framing as we know it . . . and the future of media effects. *Mass communication and society*, 19(1):7–23.
- Dallas Card, Amber E. Boydston, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. [The media frames corpus: Annotations of frames across issues](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.
- Dallas Card, Serina Chang, Chris Becker, Julia Mendelsohn, Rob Voigt, Leah Boustan, Ran Abramitzky, and Dan Jurafsky. 2022. [Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration](#). *Proceedings of the National Academy of Sciences*, 119(31):e2120510119.
- Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022. It’s not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.
- Wei-Fan Chen, Khalid Al Khatib, Benno Stein, and Henning Wachsmuth. 2020a. [Detecting media bias in news articles using Gaussian bias distributions](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4290–4300, Online. Association for Computational Linguistics.
- Wei-Fan Chen, Khalid Al Khatib, Henning Wachsmuth, and Benno Stein. 2020b. [Analyzing political bias and unfairness in news articles at different levels of granularity](#). In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 149–154, Online. Association for Computational Linguistics.
- Wei-Fan Chen, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2018. [Learning to flip the bias of news headlines](#). In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 79–88, Tilburg University, The Netherlands. Association for Computational Linguistics.
- Dennis Chong and James N. Druckman. 2007. [Framing theory](#). *Annual Review of Political Science*, 10(1):103–126.
- Kenneth Ward Church and Patrick Hanks. 1990. [Word association norms, mutual information, and lexicography](#). *Computational Linguistics*, 16(1):22–29.
- Hsiang Iris Chyi and Maxwell McCombs. 2004. [Media salience and the process of framing: Coverage of the Columbine school shootings](#). *Journalism & Mass Communication Quarterly*, 81(1):22–35.
- Alexander Dallmann, Florian Lemmerich, Daniel Zoller, and Andreas Hotho. 2015. [Media bias in German online newspapers](#). In *Proceedings of the 26th ACM Conference on Hypertext & Social Media, HT ’15*, page 133–137, New York, NY, USA. Association for Computing Machinery.
- Claes H. de Vreese. 2005. [News framing: Theory and typology](#). *Information Design Journal*, 13(1):51–62.
- Robert M. Entman. 1993. [Framing: Toward clarification of a fractured paradigm](#). *Journal of Communication*, 43(4).
- Robert M. Entman. 2007. [Framing bias: Media in the distribution of power](#). *Journal of Communication*, 57(1):163–173.
- Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. [In plain sight: Media bias through the lens of factual reporting](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6343–6349, Hong Kong, China. Association for Computational Linguistics.
- Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. [Framing and agenda-setting in Russian news: a computational analysis of intricate political strategies](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3570–3580, Brussels, Belgium. Association for Computational Linguistics.
- Lea Frermann, Jiatong Li, Shima Khanehzar, and Gosia Mikolajczak. 2023. [Conflicts, villains, resolutions: Towards models of narrative media framing](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8712–8732, Toronto, Canada. Association for Computational Linguistics.

- Matthew Gentzkow and Jesse M Shapiro. 2010. [What drives media slant? evidence from U.S. daily newspapers](#). *Econometrica*, 78(1):35–71.
- Salma Ghanem. 1997. *Filling in the Tapestry: The Second Level of Agenda Setting*, 1 edition, pages 3–14. Routledge, New York.
- David Giles and Rachel L Shaw. 2009. [The psychology of news influence and the development of media framing analysis](#). *Social and personality psychology compass*, 3(4):375–393.
- Kimberly Gross. 2008. [Framing persuasive appeals: Episodic and thematic framing, emotional response, and policy opinion](#). *Political Psychology*, 29(2):169–192.
- Shijia Guo and Kenny Q. Zhu. 2022. [Modeling multi-level context for informational bias detection by contrastive learning and sentential graph network](#). *CoRR*, abs/2201.10376.
- Felix Hamborg. 2020. [Media bias, the social sciences, and NLP: Automating frame analyses to identify bias by word choice and labeling](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 79–87, Online. Association for Computational Linguistics.
- Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. [Automated identification of media bias in news articles: An interdisciplinary literature review](#). *International Journal on Digital Libraries (IJDL)*, 20(4):391–415.
- Mareike Hartmann, Tallulah Jansen, Isabelle Augenstein, and Anders Søgaard. 2019. [Issue framing in online discussion fora](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1401–1407, Minneapolis, Minnesota. Association for Computational Linguistics.
- Miriam Hernández. 2018. ["Killed out of love": A frame analysis of domestic violence coverage in Hong Kong](#). *Violence Against Women*, 24(12):1454–1473.
- Jiwoo Hong, Yejin Cho, Jiyoung Han, Jaemin Jung, and James Thorne. 2023. [Disentangling structure and style: Political bias detection in news by inducing document hierarchy](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5664–5686, Singapore. Association for Computational Linguistics.
- Shanto Iyengar. 1991. *Is anyone responsible? how television frames political issues*. American Politics and Political Economy Series. University of Chicago Press, Chicago.
- Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. [Political ideology detection using recursive neural networks](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1113–1122, Baltimore, Maryland. Association for Computational Linguistics.
- Yangfeng Ji and Noah A. Smith. 2017. [Neural discourse structure for text categorization](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1005, Vancouver, Canada. Association for Computational Linguistics.
- Shima Khanehzar, Trevor Cohn, Gosia Mikolajczak, and Lea Frermann. 2023. [Probing power by prompting: Harnessing pre-trained language models for power connotation framing](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 873–885, Dubrovnik, Croatia. Association for Computational Linguistics.
- Shima Khanehzar, Trevor Cohn, Gosia Mikolajczak, Andrew Turpin, and Lea Frermann. 2021. [Framing unpacked: A semi-supervised interpretable multi-view model of media frames](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2154–2166, Online. Association for Computational Linguistics.
- Shima Khanehzar, Andrew Turpin, and Gosia Mikolajczak. 2019. [Modeling political framing across policy issues and contexts](#). In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 61–66, Sydney, Australia. Australasian Language Technology Association.
- Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. [SemEval-2019 task 4: Hyperpartisan news detection](#). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Vivek Kulkarni, Junting Ye, Steve Skiena, and William Yang Wang. 2018. [Multi-view models for political ideology detection of news articles](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3518–3527, Brussels, Belgium. Association for Computational Linguistics.
- Jonathan K. Kummerfeld and Dan Klein. 2013. [Error-driven analysis of challenges in coreference resolution](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 265–277, Seattle, Washington, USA. Association for Computational Linguistics.
- Haewoon Kwak, Jisun An, and Yong-Yeol Ahn. 2020. [A systematic media frame analysis of 1.5 million](#)

- new york times articles from 2000 to 2017. In *12th ACM Conference on Web Science*, page 305–314, Southampton United Kingdom. ACM.
- Nayeon Lee, Yejin Bang, Tiezheng Yu, Andrea Madotto, and Pascale Fung. 2022. **NeuS: Neutral multi-news summarization for mitigating framing bias**. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3131–3148, Seattle, United States. Association for Computational Linguistics.
- Yuanyuan Lei, Ruihong Huang, Lu Wang, and Nick Beauchamp. 2022. **Sentence-level media bias analysis informed by discourse structures**. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10040–10050, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Sora Lim, Adam Jatowt, Michael Färber, and Masatoshi Yoshikawa. 2020. **Annotating and analyzing biased sentences in news articles using crowdsourcing**. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1478–1484, Marseille, France. European Language Resources Association.
- Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. 2006. **Which side are you on? identifying perspectives at the document and sentence levels**. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 109–116, New York City. Association for Computational Linguistics.
- Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022a. **Testing the ability of language models to interpret figurative language**. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.
- Ruibo Liu, Lili Wang, Chenyan Jia, and Soroush Vosoughi. 2021. **Political depolarization of news articles using attribute-aware word embeddings**. In *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media, ICWSM 2021, held virtually, June 7-10, 2021*, pages 385–396. AAAI Press.
- Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. **Detecting frames in news headlines and its application to analyzing news framing trends surrounding U.S. gun violence**. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 504–514, Hong Kong, China. Association for Computational Linguistics.
- Yujian Liu, Xinliang Zhang, Kaijian Zou, Ruihong Huang, Nicholas Beauchamp, and Lu Wang. 2023. **All things considered: Detecting partisan events from news media with cross-article comparison**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15472–15488, Singapore. Association for Computational Linguistics.
- Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nicholas Beauchamp, and Lu Wang. 2022b. **POLITICS: Pretraining with same-story article comparison for ideology prediction and stance detection**. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1354–1374, Seattle, United States. Association for Computational Linguistics.
- John McCarthy, Larissa Titarenko, Clark McPhail, Patrick Rafail, and Boguslaw Augustyn. 2008. **Assessing stability in the patterns of selection bias in newspaper coverage of protest during the transition from communism in Belarus**. *Mobilization: An International Quarterly*, 13(2):127–146.
- Douglas M McLeod, Hyesun Choung, Su Min-Hsin, Kim Sang-Jung, Ran Tao, Jiawei Liu, and ByungGu Lee. 2022. **Navigating a diverse paradigm: A conceptual framework for experimental framing effects research**. *Review of communication research*, 10.
- Denis McQuail and Mark Deuze. 2020. *Mcquail's media and mass communication theory*, 7th ed edition. SAGE Publications, Thousand Oaks.
- Julia Mendelsohn, Ceren Budak, and David Jurgens. 2021. **Modeling framing in immigration discourse on social media**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2219–2263, Online. Association for Computational Linguistics.
- Stefano Menini, Federico Nanni, Simone Paolo Ponzetto, and Sara Tonelli. 2017. **Topic-based agreement and disagreement in US electoral manifestos**. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2938–2944, Copenhagen, Denmark. Association for Computational Linguistics.
- Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. **Model cards for model reporting**. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.
- Glenn W. Muschert and Dawn Carr. 2006. **Media salience and frame changing across events: Coverage of nine school shootings, 1997–2001**. *Journalism & Mass Communication Quarterly*, 83(4):747–766.
- Nona Naderi and Graeme Hirst. 2017. **Classifying frames at the sentence level in news articles**. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 536–542, Varna, Bulgaria. INCOMA Ltd.

- Sylvain Parasio. 2022. *Computing the news : data journalism and the search for objectivity*. Columbia University Press New York, New York.
- Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023a. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.
- Jakub Piskorski, Nicolas Stefanovitch, Nikolaos Nikolaidis, Giovanni Da San Martino, and Preslav Nakov. 2023b. Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3001–3022, Toronto, Canada. Association for Computational Linguistics.
- Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659, Sofia, Bulgaria. Association for Computational Linguistics.
- Douglas A. Reynolds. 2009. Gaussian mixture models. In Stan Z. Li and Anil K. Jain, editors, *Encyclopedia of Biometrics*, pages 659–663. Springer US.
- Shamik Roy and Dan Goldwasser. 2020. Weakly supervised learning of nuanced frames for analyzing polarization in news media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7698–7716, Online. Association for Computational Linguistics.
- Dietram A. Scheufele. 1999. Framing as a theory of media effects. *Journal of Communication*, 49(1):103–122.
- Timo Spinde, Felix Hamborg, Lada Rudnitskaia, and Bela Gipp. 2021a. Identification of biased terms in news articles by comparison of outlet-specific word embeddings. In *Proceedings of the iConference 2021*.
- Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021b. Neural media bias detection using distant supervision with BABE - bias annotations by experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1166–1177, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Isidora Tourni, Lei Guo, Taufiq Husada Daryanto, Fabian Zhafransyah, Edward Edberg Halim, Mona Jalal, Boqi Chen, Sha Lai, Hengchang Hu, Margrit Betke, Prakash Ishwar, and Derry Tanti Wijaya. 2021. Detecting frames in news headlines and lead images in U.S. gun violence coverage. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4037–4050, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Oren Tsur, Dan Calacci, and David Lazer. 2015. A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1629–1638, Beijing, China. Association for Computational Linguistics.
- David Vilar, Jia Xu, Luis Fernando D’Haro, and Hermann Ney. 2006. Error analysis of statistical machine translation output. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06)*, Genoa, Italy. European Language Resources Association (ELRA).
- Alden Williams. 1975. Unbiased study of television news bias. *Journal of Communication*, 25(4):190–199.
- Caleb Ziems and Diyi Yang. 2021. To protect and to serve? analyzing entity-centric framing of police violence. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 957–976, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A List of Papers Included

Table 2 (on the next page) lists our body of literature, identified as described in Section 1.1, and indicates which of our three disconnects are addressed in each paper (if any). The table caption explains our labelling procedure.

Paper	Local/Global	Dynamics	Comparison
Ajjour et al. (2019)			
Aksenov et al. (2021)			
Akyürek et al. (2020)			
Ali and Hassan (2022)			
Baly et al. (2020)			
Baumer et al. (2015)			
Cacciatore et al. (2016)			
Card et al. (2015)			
Card et al. (2022)		x	
Chen et al. (2020a)	x		
Chen et al. (2020b)	x		
Chen et al. (2018)			
Chong and Druckman (2007)			
Chyi and McCombs (2004)			
Dallmann et al. (2015)			
de Vreese (2005)			
Entman (1993)			
Entman (2007)			
Fan et al. (2019)	x		
Field et al. (2018)		x	
Frermann et al. (2023)	x		
Gentzkow and Shapiro (2010)			
Ghanem (1997)			
Giles and Shaw (2009)			
Gross (2008)			
Guo and Zhu (2022)			
Hamborg (2020)			
Hamborg et al. (2019)			
Hartmann et al. (2019)			
Hernández (2018)	x	x	
Hong et al. (2023)	x		
Iyyer et al. (2014)			
Ji and Smith (2017)			
Khanehzar et al. (2023)			
Khanehzar et al. (2019)			
Khanehzar et al. (2021)			
Kiesel et al. (2019)			
Kulkarni et al. (2018)			
Kwak et al. (2020)		x	

Continued on next page

Paper	Local/Global	Dynamics	Comparison
Lee et al. (2022)			
Lei et al. (2022)			
Lim et al. (2020)			
Lin et al. (2006)			
Liu et al. (2021)			
Liu et al. (2019)			
Liu et al. (2023)			x
McCarthy et al. (2008)	x		
McLeod et al. (2022)			
McQuail and Deuze (2020)			
Mendelsohn et al. (2021)			
Menini et al. (2017)			
Muschert and Carr (2006)	x		
Naderi and Hirst (2017)			
Piskorski et al. (2023b)			
Recasens et al. (2013)			
Roy and Goldwasser (2020)			
Scheufele (1999)			
Spinde et al. (2021a)			
Spinde et al. (2021b)			
Tourni et al. (2021)			
Tsur et al. (2015)			
Williams (1975)	x	x	
Ziems and Yang (2021)		x	
Total	9	6	1

Table 2: Cited Literature. Papers marked as ‘Local/Global’ analyse media bias or framing, or provide data at different levels of granularity, ranging from words and sentences (or spans) to entire documents. For a paper to consider ‘Dynamics’, we required the study to include an analysis of the development of a topic across a specific axis, either temporal or spatial (across countries). Papers marked in the ‘Comparison’ column characterise bias or framing by explicitly contrasting data samples from different ideologies or political leanings.

The Crime of Being Poor: Associations between Crime and Poverty on Social Media in Eight Countries

Georgina Curto,¹ Svetlana Kiritchenko,² Kathleen C. Fraser,² and Isar Nejadgholi²

¹University of Notre Dame, Notre Dame, USA

²National Research Council Canada, Ottawa, Canada

gcurtore@nd.edu, {svetlana.kiritchenko,kathleen.fraser,isar.nejadgholi}@nrc-cnrc.gc.ca

Abstract

Content Warning: This paper presents textual examples that may be offensive or upsetting.

Negative public perceptions of people living in poverty can hamper policies and programs that aim to help the poor. One prominent example of social bias and discrimination against people in need is the persistent *association of poverty with criminality*. The phenomenon has two facets: first, the belief that poor people are more likely to engage in crime (e.g., stealing, mugging, violence) and second, the view that certain behaviors directly resulting from poverty (e.g., living outside, panhandling) warrant criminal punishment. In this paper, we use large language models (LLMs) to identify examples of crime–poverty association (CPA) in English social media texts. We analyze the online discourse on CPA across eight geographically diverse countries, and find evidence that the CPA rates are higher within the sample obtained from the U.S. and Canada, as compared to the other countries such as South Africa, despite the latter having higher poverty, criminality, and inequality indexes. We further uncover and analyze the most common themes in CPA posts and find more negative and biased attitudes toward people living in poverty in posts from the U.S. and Canada. These results could partially be explained by cultural factors related to the tendency to overestimate the equality of opportunities and social mobility in the U.S. and Canada. These findings have consequences for policy-making and open a new path of research for poverty mitigation with the focus not only on the redistribution of wealth but also on the mitigation of bias and discrimination against people in need.

1 Introduction

Computational methods provide new insights that can trigger innovative interventions towards the UN Sustainable Development Goals (SDGs) (Vinueza et al., 2020). The “end of poverty in all its

forms everywhere” is the #1 UN SDG and constitutes an urgent call for action. However, there are still 685M people living in extreme poverty worldwide (The World Bank, 2023), and COVID-19 has particularly affected the poorest: the number of people living in extreme poverty rose by 11% in 2020 (The World Bank, 2022). Poverty affects not only the population in developing regions but also a significant percentage of those living in thriving economies (Creamer et al., 2022; Eurostat, 2022): in the United States, 11.6% of the population (37.9M people) are in a situation of poverty (Creamer et al., 2022), and within the EU-27, there are 95.3M people (22% of the population) at risk of poverty (Eurostat, 2022). In this context, innovative measures are required to work towards poverty mitigation across the globe. Traditional policies based on the redistribution of wealth may not be enough, as evidenced by a deceleration in the poverty reduction rates in the last decades (Claudia et al., 2018). Interdisciplinary research, incorporating AI-enabled tools into these efforts, offers perspectives previously unavailable with traditional quantitative and qualitative approaches.

Prejudice against people with low socioeconomic status can hinder poverty reduction efforts (Arneson, 1997; Everatt, 2009). When the poor are believed to be responsible for their situation and, therefore, unworthy of help (“undeserving poor”), it is difficult for policy makers to approve and implement poverty reduction policies (Nunn and Biressi, 2009). Therefore, the blameworthiness of the poor could have an impact on the actual poverty levels. However, aporophobia—a social bias of rejection and contempt for the poor and the associated institutional discrimination of the group (Cortina, 2022)—has only recently become a topic of research, with initial studies providing empirical evidence of this phenomenon (Curto et al., 2022).

An important facet of aporophobia is a frequent *association of poverty with criminality* in society

(Kiritchenko et al., 2023). It can be seen (1) in public opinion as individuals’ beliefs and stereotypes (we refer to this as *crime–poverty bias*), and (2) in discriminatory actions at an institutional level (we refer to it as *institutional criminalization of the poor*). In the first instance, homeless people are stereotyped as threatening, violent, not wanting to work, and mentally ill (Faragó et al., 2022), and poor people, as a group, are believed to be frequently involved in criminal activities, such as theft and illegal drug dealing. In turn, the criminal offenses devised for sleeping rough in many cities of the so-called developed countries are an example of the institutional criminalization of the poor (Barcelona City Council, 2005). Other examples include probation and even incarceration for people who cannot afford to pay minor fines (Geraghty, 2015; Terradillos Basoco, 2020).

The study of the crime–poverty association (CPA) needs to be rooted in cognitive science and the philosophy of discrimination. Allport (1954) explains that human beings interpret information by classifying it into categories based on their previous experience. This process is at the origin of prejudices, which have been described as overgeneralized and, therefore, misleading beliefs that result in systematic and predictable errors in decision making based on available heuristics (Kahneman, 2011). Prejudices can lead to social bias reflected in derogatory speech (Ely, 1980; Greenwald et al., 2003), and even to discrimination through unfair actions towards members of a group (Allport, 1954).

In this paper, we use NLP techniques to analyze the prevalence of social media discourse around CPA in eight geographically-diverse countries where English is an official or majority language. Despite the fact that social networks are not representative of the whole population, they do act as a partial mirror that allow researchers to measure and track societal biases and discriminatory behaviors that might be hard to detect by traditional quantitative analyses, such as surveys. We employ pre-trained large language models (LLMs) to classify 500K social media posts in English from the selected countries, identifying statements discussing CPA. We further conduct a topic-modelling analysis to examine the *content* of the retrieved posts and determine which aspects of CPA are more salient in different regions. Finally, we contextualize our findings by comparing them with published indicators of poverty, criminality, and inequality in each of the countries in our study. The preliminary

results offered in this article inform about the differences in the social discourse around CPA within the countries in scope and open lines of research towards the mitigation of poverty by acting on biases and discriminatory actions that affect people in need.

2 Related Work

We briefly review some of the related social science research on the correlation between crime and poverty, as well as computational work on detecting and responding to social biases.

2.1 Correlation between Poverty and Criminality

Previous research provides evidence of the correlation between poverty and criminality (Looney and Turner, 2018; Becker, 1968). For example, economists have examined the correlation between poverty and property crimes, assessing the cost-benefit analysis (Freeman, 1999; Wu and Wu, 2012; Costantini et al., 2018). Research in criminology also correlates poverty and violent crimes, explained by the strain generated by individuals’ failure to achieve socially valued objectives (Agnew, 1992, 2001). Within the context of the U.S., studies found that children who grew up in families in the bottom 10% of the income distribution are 20 times more likely to go to prison in their early 30s than children born in top-decile families. Further, one in ten boys born to families with the lowest income decile are in prison at age 30, and they account for 27% of prisoners of that age (Looney and Turner, 2018). Similarly, other multidimensional factors associated with poverty, such as social determinants of health and education, appear to have an impact on incarceration (Miller, 2013; Hinton, 2017). In turn, the mark of a criminal record generates impediments for employment (Pager, 2003; Mueller-Smith, 2015), which constitutes a vicious circle to get out of poverty. Despite the fact that poverty and criminality are correlated, the strength of the correlation can be greatly over-estimated in public opinion, leading to bias and discrimination against people in need. In this paper, we examine online discussions on crime–poverty association (in the form of bias and discrimination) and explore socio-economic factors as well as cultural differences that might affect the prevalence of CPA discussions in certain regions of the world.

It must be highlighted that the existing corre-

lation between poverty and criminality does not justify accepting stereotypes and acts of discrimination that associate people in need, as a group, with crime. As is the case with gender discrimination, racism or xenophobia, this type of shared generalization exacerbates underlying social biases and generates a vicious circle for vulnerable social groups. Further, generalizations that associate people in poverty and crime are detrimental to the dignity of the persons affected, and thwart the efforts towards poverty reduction. Bearing in mind the urgency to alleviate poverty, this type of bias and discrimination needs to be identified, tracked and mitigated.

2.2 Addressing Social Bias with NLP

A significant research effort in NLP has been dedicated to identification and mitigation of social bias in human-written text, particularly in social media. This includes works on stereotype detection (Fokkens et al., 2018; Marzouki et al., 2020; Charlesworth et al., 2021; Fraser et al., 2022) and identifying and countering hate speech and toxic language (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018; Vidgen et al., 2019; Tekiroğlu et al., 2020; Kiritchenko et al., 2021; Zhu and Bhat, 2021; Ashida and Komachi, 2022). However, previous work mostly focused on identifying sub-types of harmful language that target specific groups, such as sexism (Istaiteh et al., 2020; Chiril et al., 2020; Samory et al., 2021), racism (Istaiteh et al., 2020; Waseem, 2016; Davidson et al., 2017), and anti-immigrant hoaxes (Bourgeade et al., 2023).

Only a handful of studies have investigated bias based on socio-economic status. Curto et al. (2022) provided evidence of bias against poor people in word embeddings, built on Google News, Twitter, and Wikipedia corpora. Perez Almendros et al. (2020) collected and annotated a dataset with patronizing and condescending language that included homeless and poor people among the selected target groups. Kiritchenko et al. (2023) argued that existing toxic language detection models and datasets are inadequate to effectively identify bias against people with low socio-economic status. We continue and extend this line of work to analyze the social media discourse on social bias and institutional criminalization of the poor in different regions of the world.

3 Data and Methods

In this project, we aim to analyze and compare the social media discourse on CPA, expressed both in the form of bias and institutional criminalization of the poor, in various regions of the world. We choose Twitter as our data source, now called X,¹ which provides a rich stream of everyday conversations of ordinary people on a variety of topics. We start by collecting a large set of tweets in English referring to poor people, written by users from eight geographically-diverse countries (Section 3.1).

Next, we automatically retrieve tweet sentences that refer to an association between criminality and poverty. We approach this task as binary text classification with two categories: ‘text refers to a crime–poverty association’ (or ‘CPA’) and ‘no reference to a crime–poverty association’ (or ‘not CPA’). This task has not been extensively studied in NLP, and there are no annotated data available. Thus, we turn to pre-trained large language models as zero-shot or few-shot text classifiers (Wang et al., 2023). While text classifiers, in general, benefit from fine-tuning on task-specific labelled datasets, LLMs have shown to be effective classifiers in scenarios with limited resources (Chae and Davidson, 2023). This is specifically relevant to our study, where we analyze various aspects of CPA in social media discourse across different countries at an aggregate level, and individual labels assigned by LLMs are not critical. For model selection and validation, we collect and manually label a small test set (described in Section 3.2), on which we evaluate the performance of several state-of-the-art LLMs, listed in Section 3.3.

After identifying the best performing LLM on the test set, we prompt it to automatically categorize the full set of tweet sentences into the ‘CPA’ and ‘not CPA’ categories. Using this approach, we compute and compare the percentages of tweets about poor people that refer to an association between criminality and poverty, in the eight countries. We further examine the CPA posts to discover and analyze common themes on the issues of criminality and poverty around the world using unsupervised topic modeling (Section 3.4). To contextualize our findings, we compare them with the published statistics on various economic and criminality indicators in the studied countries, and

¹Since the platform was called Twitter when we collected the data, we use that name throughout. Note that we collected the data prior to the introduction of the paywall.

speculate on possible reasons for the observed discrepancies in the Discussion (Section 5). In the following, we describe the dataset and methodology in detail.

3.1 Twitter Corpus

We use the Twitter Research API to collect English tweets pertaining to poor people from 25 August 2022 to 23 November 2022. We first collected a set of query terms from social psychology literature and augmented this set with synonyms and related terms. We then collected a one-week sample of tweets using this set of query terms and manually examined the retrieved tweets to discard the terms that resulted in very small numbers of retrieved tweets or many irrelevant tweets. The final list of query terms includes: *the poor* (used as a noun as opposed to an adjective, as in ‘the poor performance’), *poor people*, *poor ppl*, *poor folks*, *poor families*, *homeless*, *on welfare*, *welfare recipients*, *low-income*, *underprivileged*, *disadvantaged*, *lower class*. The single word *poor* is not included as a query term because of its polysemy (it can apply to people, but can also be used to describe other things, e.g., ‘poor results’). We exclude explicitly offensive terms that tend to be used in personal insults, such as *trailer trash* or *scrounger*.

Re-tweets, tweets with more than five hashtags, and tweets with URLs are excluded. Tweets written by bots (identified as user accounts with the user or screen name including the word ‘bot’) are excluded as well. This filtering step helps remove posts from commercial accounts. Since tweets can be up to 280 characters and include several sentences, we split each tweet into individual sentences and keep only sentences that include at least one of the query terms. In total, there are over 1.3 million sentences in the corpus.

We are also interested in the geographical locations from which tweets originated. Unfortunately, only about 2% of tweets include the exact geographical information. Therefore, in addition to the tweet location (‘place’ field), we rely on user location that users voluntarily provide in their Twitter accounts, which is available for about 60% of posts. The user location is recorded as a free-form text, and tweeters are often very creative in describing their location (e.g., “somewhere on Earth”). We automatically parse user location descriptions to extract country information for the most frequently mentioned countries. In the absence of a country name, we consider the mentions of U.S. states,

Location	# of sentences
United States	326,993
United Kingdom	80,947
Canada	32,978
India	14,029
Nigeria	10,529
Australia	9,698
South Africa	7,729
Kenya	3,378
<hr/>	
Total (eight countries)	486,281
<hr/>	
Other locations	337,252
No location information	539,365
<hr/>	
Total	1,362,898

Table 1: The number of sentences in the tweet corpus per geographical location.

Canadian provinces, and major cities in the U.S., U.K. and Canada, since these are also commonly used by tweeters. (Major cities from other countries are rarely used without the country name.) Table 1 shows the number of sentences in the corpus per geographical location. In the following analysis, we focus on eight geographically-diverse countries with English as an official or majority language, for which the corpus contains at least 1,000 sentences: the United States of America, the United Kingdom, Canada, India, Nigeria, Australia, South Africa, and Kenya. There are, in total, 486,281 sentences written by tweeters from these eight countries.

3.2 Manually Labeled Evaluation Datasets

To validate and test automatic LLM-based classifiers, we manually annotate a small portion of the dataset, identifying examples of CPA. We capture both statements that illustrate ‘crime-poverty bias’ expressed in public opinion and sentences describing ‘institutional criminalization of the poor’. In particular, we label as ‘CPA’:

1. ‘Crime-poverty bias’: utterances that state or refer to the belief that people living in poverty, as a group, are more likely to be engaged in illegal activities (e.g., stealing, harassing other people, selling illegal drugs, etc.);
2. ‘Institutional criminalization of the poor’: utterances that state that poor people are more likely to face legal consequences due to their lack of financial resources (e.g., being put and kept in jail for minor offenses or unpaid fees/debts, inability to pay bail or hire lawyers, anti-homelessness laws and actions, etc.).

Both categories characterize CPA and need to be

CPA	Examples of statements illustrating ‘crime-poverty bias’:
	Poor people just steal other people’s property rather than get a job, earn money and buy what they need. Urban inner-city crime has been real since we had cities filled with poor people. It’s been like this since forever so you can say it is normal for poor people to commit crimes.
	Examples of statements illustrating ‘institutional criminalization of the poor’: Law enforcement and prisons are routinely used against poor people not because of safety, but to protect the wealthy. Poor people from poor neighborhoods are easy targets for incarceration, police extortion and police murders. Cities do have bylaws to discard the homeless like trash.
Not CPA	Examples of statements not expressing an association between poverty and criminality:
	Democratic states are overrun with crime, homelessness, drugs, and debt. They haven’t stopped the crime and do little for the homeless. The government is stealing from the poor to give to the rich.

Table 2: Example sentences referring to an association of poverty with criminality (top), and examples that are not labelled as making the association (bottom). All tweet sentences are paraphrased to protect the privacy of the users.

Dataset	Total sent.	CPA sent. (% of total)
Test ‘Random’	400	12 (3%)
Test ‘Crime’	400	151 (38%)
Development	107	38 (36%)
Total	907	201 (22%)

Table 3: The number of sentences in each manually labeled test set.

examined side by side. Moreover, each one supports and reinforces the other. Harmful beliefs and stereotypes perpetuated in online communications affect the overall public perception of the group and build a strong foundation for systemic policies. On the other hand, disproportional police surveillance and incarcerations of poor people lead to statistical evidence supporting the stereotypes.

Note that the CPA category does not only include utterances that *perpetuate* the belief of poor people being criminals or *support* the laws and policies discriminating against the poor. In fact, many writers criticize such beliefs and advocate for stronger social support and better policies that would empower people in poverty and help them overcome the adversities. Although such posts do not directly contribute to reinforcing the crime–poverty association, we also label such messages with the CPA category because they are evidence of the existence of the phenomenon.

We first randomly select 50 sentences originating from each of the eight countries of interest (400 sentences in total) and annotate them for CPA. We call this test set ‘Random’. Only 12 sentences (3%) in this set are labeled as CPA. To increase the proportion of CPA messages, we create a second test set, called ‘Crime’, using the following data enrichment procedure. We first select sentences

Dataset	% agree	Cohen’s κ
Test ‘Random’	97.0	0.49
Test ‘Crime’	88.0	0.74
Development	87.9	0.73

Table 4: Inter-annotator agreement on each test set.

from the tweet corpus that include at least one of the following crime-related words: *crime*, *crimes*, *criminal*², *jail**, *prison**, *arrest**, *police*, *cops*, *policing*, *imprison**, *incarcerat**, *prosecut**, *assault**, *harass**, *steal*, *stealing*, *stolen*, *stole*, *theft*. From these sentences, we again randomly select 50 sentences originating from each of the eight countries of interest (400 sentences in total) and annotate them for CPA. In this test set, 151 sentences (38%) are labeled as CPA. Table 2 shows example sentences for both crime–poverty bias in public opinion and institutional criminalization of the poor (labeled as CPA), and sentences that mention crime but are not labelled as CPA since they do not imply bias or discrimination against the group. For the latter, in many cases, crime and poverty are both mentioned as problems that need to be addressed but are not causally related. In other cases, poor people are described as the victims, rather than the perpetrators, of crime.

Finally, a smaller development set was created in a similar way as the test set ‘Crime’. Table 3 shows the details of the three datasets.

Two authors of this paper annotated the sentences independently, and then all disagreements were discussed and resolved. Table 4 shows the inter-annotator agreement on the three evaluation

²The wildcard * indicates any number of alphabetic characters to cover morphological variants of the word (e.g., criminal, criminals, criminalization, etc.)

datasets. The agreement is measured in two ways: (1) as the percentage of sentences on which the two annotators agree on the label, and (2) as Cohen’s κ . Both metrics demonstrate moderate to substantial levels of agreement.

3.3 Automatic Classification of CPA

After manually annotating 907 sentences in the previous section, we identified 201 instances of CPA. Since this is not sufficient to train a classifier to detect CPA from scratch, we instead investigate the possibility of using methods not requiring large amounts of data for training, such as zero- or few-shot learning with LLMs. We use the small annotated dataset to evaluate and compare the models.

We experiment with two open-source LLMs and one commercial model:

1. **Llama 2** (7B), an open-source model released by Meta (Touvron et al., 2023)
2. **Flan-T5** (XL), an open-source model created by Google (Chung et al., 2022)
3. **ChatGPT** (GPT-3.5-turbo-0125), a commercial model produced by OpenAI.

We prompt the LLMs in both zero-shot and few-shot settings. It has been shown that LLMs’ performance varies significantly with even minor variations in prompts, and prompts that are optimal for one model might not perform well for another model (Voronov et al., 2024). Therefore, we design zero-shot and few-shot prompts for each of the models separately. The best prompts for each model are reported in Appendix A. Temperature was set to zero for all three LLMs.

3.4 Unsupervised Topic Modeling

Next, we examine the content of CPA posts to discover the most salient themes in the discussions on the association between criminality and poverty in the different regions. We apply unsupervised topic modeling on all the tweet sentences automatically classified as CPA. Topic modeling is a helpful tool to quickly analyze semantic content of large amounts of text. We use BERTopic (Grootendorst, 2022), a flexible state-of-the-art toolkit for unsupervised, semi-supervised, and supervised topic modeling. BERTopic employs a density-based clustering technique HDBSCAN (Campello et al., 2013), which identifies dense regions in the text representation space and leaves texts outside these dense

regions as outliers. These dense regions would represent the most commonly discussed topics in tweets associating poor people and criminality.

After a few preliminary experiments, we set the parameters as follows. Texts are converted to numerical vectors using sentence transformers³ with the *roberta-large-nli-mean-tokens* pre-trained embedding model. We use CountVectorizer⁴ as the vectorizer model, remove stopwords and terms that appear in less than 5% of the sentences ($min_df = 0.05$), and set the minimum size of the clusters as $min_cluster_size = 100$. For all the other parameters, the default settings of the BERTopic package are used.

We analyze the most common topics discovered by BERTopic and compare their prevalence in Twitter discussions in the eight countries in scope.

4 Results

First, we report the results obtained from the automatic LLM-based classification of a large collection of tweets to identify CPA for the eight countries in scope (Section 4.1). We analyze these results in the context of economic and criminality indicators for each country. Next, we look at the content of CPA posts and examine common themes in the Twitter discourse around CPA in different parts of the world (Section 4.2).

4.1 Crime–Poverty Association in Tweets

We evaluate the performance of the three LLMs with their respective best zero- and few-shot prompts on the development and test sets. Table 5 reports the results. Among the three tested LLMs, on all three datasets, the best performance is achieved by ChatGPT (GPT-3.5-turbo) using a few-shot prompt. Therefore, we proceed with that model in the following analysis of CPA in different regions of the world.

We prompt ChatGPT with the selected few-shot prompt on all 486,281 sentences of the original corpus written by users from the eight countries in scope. In total, 38,034 sentences (8%) were classified by ChatGPT as CPA. Table 6 shows the percentages of CPA sentences per country.

The results indicate that tweeters from the United States and Canada are more likely to re-

³<https://github.com/UKPLab/sentence-transformers>

⁴https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

Classification method	Dev. set			Test set 'Crime'			Test set 'Random'		
	Acc.	F_{avg}	F_{CPA}	Acc.	F_{avg}	F_{CPA}	Acc.	F_{avg}	F_{CPA}
Llama 2									
zero-shot	0.76	0.75	0.61	0.71	0.70	0.57	0.96	0.96	0.38
few-shot	0.77	0.77	0.71	0.69	0.70	0.61	0.94	0.95	0.40
Flan-T5									
zero-shot	0.74	0.74	0.62	0.70	0.71	0.62	0.95	0.96	0.36
few-shot	0.76	0.72	0.50	0.73	0.69	0.49	0.97	0.97	0.44
GPT-3.5-turbo									
zero-shot	0.79	0.78	0.63	0.71	0.70	0.57	0.97	0.97	0.40
few-shot	0.83	0.83	0.76	0.77	0.77	0.69	0.97	0.98	0.67

Table 5: Performance of the three LLMs on the evaluation datasets: accuracy (Acc.), support-weighted macro-averaged F1-score (F_{avg}), and F1-score on the CPA category (F_{CPA}). Highest values in each column are in bold.

fer to and discuss crime–poverty associations in their posts than users in the other countries. To contextualize these findings, Figure 1 presents the percentage of sentences labeled as CPA for each of the eight countries together with their economic and crime indicators: overall criminality rate, poverty headcount ratio at \$2.5 a day (purchasing power adjusted prices), inequality indicators (Gini Index and 10% income share), and unemployment rate. The results show that the United States and Canada have the highest CPA, a sign of pervasiveness of the phenomenon in these countries, despite having comparable or even lower poverty, criminality, inequality and unemployment rates than the other countries in scope. It is also worth noting that India, which has a higher poverty headcount and similar levels of inequality and unemployment rates to the United States, has the lowest CPA rate, which may indicate that crime and poverty are seen as less associated with each other and discussed as separate negative factors in this society. Also, online posts from South Africa discuss CPA at a substantially lower rate than in the U.S. and Canada, despite South Africa having the highest levels of inequality, criminality, and unemployment among the countries included in the study. We discuss factors that might influence CPA, in addition to socioeconomic indicators, in Section 5.

4.2 Common Themes in CPA Tweets

To examine the content of Twitter discussions related to crime–poverty associations, we apply unsupervised topic modeling using BERTopic on the 38,034 sentences labeled by ChatGPT as CPA. BERTopic extracts 20 topics (ordered by the number of sentences in a topic), and leaves around 55% of sentences unclustered. A few topics are grouped by the targeted subpopulation, e.g., *homeless people* or *people on welfare*, but semantically represent

a wide mix of themes pertaining to the group. One such topic to note is Topic 11, which discusses the prevalence of Black people amongst the economically disadvantaged, racial discrimination aggravated by aporophobic attitudes, and other topics on the intersection of race and economic inequality. However, we exclude such topics from our current analysis as we aim to focus on themes relating to criminality in general, and not to specific subgroups.

Of the 20 initial topics, we select ten most interpretable topics that could be mapped to a social theme.⁵ Figure 2 shows how often these ten topics are discussed in each of the eight countries (i.e., the shading in cell (i, j) represents the proportion of tweet sentences from Country $_j$ that are clustered in Topic $_i$).

The topic modeling analysis reveals major differences in the social media discourse around poverty in North America as opposed to the other examined countries. The highest proportions of sentences from the U.S. and Canada fall in Topics 1, 3, and 4. Topic 1 comprises sentences expressing negative attitudes towards homeless people, portraying them as criminals and drug addicts (e.g., “*homeless crime is a huge issue*”, “*rampant homeless addicts make it an unsafe place to live*”, “*filth and petty crimes from homeless*”). Topic 3 includes sentences on homelessness being illegal (e.g., “*if you’re homeless you are illegal*”), and Topic 4 consists of calls for authorities to remove homeless people from the streets (e.g., “*get the homeless, druggies off the streets*”, “*clear away homeless en-*

⁵Note that the topic interpretation and mapping to social themes are done manually by the authors and, therefore, subjective. Not all sentences in a topic discuss the corresponding theme in the same way, and some topics may contain opposing views. The manual label assigned to a topic is intended to represent the interpretation for the majority of sentences in that topic.

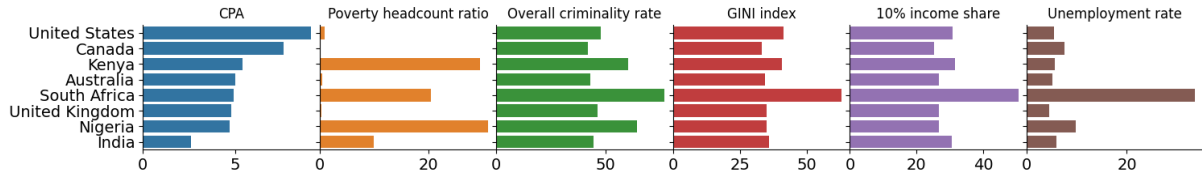


Figure 1: For the countries in scope, percentage of CPA discussions in tweets related to poor people, poverty headcount at \$2.5 a day (purchasing power adjusted prices), overall criminality rate, indicators of inequality (Gini index and 10% income share), and unemployment rates. Sources: poverty headcount ratio, Gini index, and 10% income share rates are as reported by the World Bank (2017 or nearest year); unemployment rates are from the World Bank (2021); overall criminality rates are from worldpopulationreview.com.

Country	% CPA
United States	9.1
Canada	7.6
Kenya	5.4
Australia	5.0
South Africa	4.9
United Kingdom	4.8
Nigeria	4.7
India	2.6

Table 6: Percentage of sentences from each country classified as CPA by ChatGPT.

campments”). Further, in Topic 7 tweeters from the U.S. and Canada talk about various incidents of homeless people involved in violent attacks and other crimes (e.g., “a person was mugged by a homeless dude with an axe”). Other countries include substantially less posts in these topics.

In contrast, Topics 5 and 6 are among the largest topics for the other countries, especially for Nigeria, Kenya, and India. These topics comprise posts describing the discriminative nature of laws and policies disproportionately affecting people with less resources and power (e.g., “laws are made for the poor”, “rules only affect the poor”). A significant number of posts from the U.K. also appear in Topics 12, 14, and 16, where tweeters criticize government policies that aggravate the hardship of living in poverty or punish poor people for being poor (e.g., “government continues persecution of the poor”, “a tool with which to mass murder the poor, and then blame them for it”). Finally, in Topic 9, tweeters from all countries denounce the current state of social structure where poor and homeless people are frequently arrested and kept in jails for minor offenses or no reason at all (e.g., “meanwhile, poor people guilty or not guilty of less substantial crimes are going straight to prison”).

5 Discussion

Our results reveal substantial quantitative and thematic differences in online discourse related to the crime–poverty association in the eight studied countries. The proportion of CPA discussions in Twitter in the U.S. is almost twice as high as in the other countries in scope, both in the Global North and Global South. Furthermore, posts from the U.S. and Canada more often express bias against poor people as a group, associating them with different kinds of crime. There are many posts portraying homeless people as thieves, muggers, and drug addicts, and calling for systemic measures to remove them from the streets. People receiving welfare support are often described as scammers and not wanting to work. In contrast, tweeters from the other countries are more concerned with unfair laws and regulations that disproportionately affect the poor and aggravate their hardship.

A potential explanation for these findings could be found in the narrative shared by the United States and Canada of being the “lands of opportunity”, where the rich and the poor are thought to have equal chances for success (United Nations, 2018). The poor, therefore, would be blamed and even punished for their inability to get out of poverty (Desmond, 2023). However, the principle of equal opportunity can be considered an oxymoron since every person is exposed to different opportunities in life from the moment of birth (Sandel, 2020), and the job market for individuals with low educational qualifications, disability, and with no assistance to find employment is very limited. The indicators of social mobility and inequality support the claim from the United Nations that the poor in the United States are overwhelmingly those born into poverty (United Nations, 2018): intergenerational social mobility in the United States from the bottom to the top income quantile is as

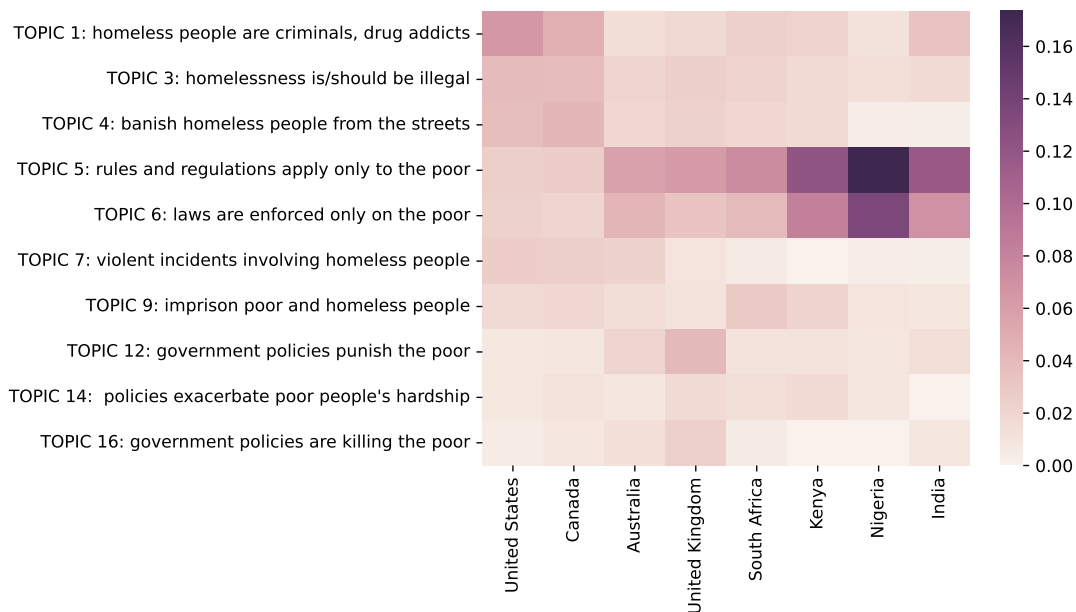


Figure 2: The proportion of tweet sentences from each country in ten most interpretable topics.

low as 7.8%, below European countries such as the U.K., France, Italy, or Sweden (Alesina et al., 2018). In fact, intergenerational mobility has declined substantially over the last 150 years in the United States (Song et al., 2020) and income inequality has been growing since the 1980s (The World Bank, 2023).

6 Conclusions

This paper presents a quantitative and a thematic analyses of the prevalence of online discussions on the association of criminality and poverty in eight geographically-diverse countries. It provides evidence that such discussions more frequently occur within the U.S. and Canada, as compared to the other countries such as South Africa, despite the latter having higher poverty, criminality, and inequality indexes. Moreover, the most prevalent topics in these discussions in the U.S. and Canada demonstrate negative attitudes and social bias against people in need, for example, portraying homeless people as criminals, muggers, and drug addicts. In contrast, in the other countries in scope, such as Nigeria, India or Kenya, the most frequently discussed topics refer to the unfairness of laws that do not provide support or even discriminate against the poor. We speculate that these differences can be partially attributed to the rhetoric of equal opportunities widespread in the U.S. and Canada. The insights obtained from this study shed light towards a new path of research for poverty mitigation, where

the focus should be not only on the redistribution of wealth but also on the mitigation of social bias and discrimination against the poor.

Limitations

This article constitutes a preliminary analysis of CPA in social media discourse, including both crime–poverty bias in public opinion and institutional criminalization of the poor. As such, it offers evidence of the phenomenon and aims to open a new line of research that needs to go deeper and explore the different types of stereotypes and acts of discrimination that contribute to the phenomenon.

The presented analysis covers only English-language posts from eight countries representing different regions of the world. While English is an official or majority language in the selected countries, other languages are also widely spoken in some of these countries. The user posts are collected using a pre-specified set of terms in standard English that may exclude related terms in regional dialects. Further, the posts are collected only from one social media platform, Twitter. Similar to any other social network, Twitter represents a non-random sample of the general population in terms of gender, age, ethnicity, and other socio-demographic characteristics (Mislove et al., 2011). In particular, it is predominantly used in the United States (Barbieri et al., 2020). Therefore, the findings from this study may not generalize to population at large. Future work will extend this analysis

to cover other languages, geographical locations, and sources of public opinion. In addition, the inclusion of multi-modal data, combining text with images and video, can enrich the analysis of social media discourse on poverty and criminality.

The analysis relies on automatic methods of data collection and categorization. While allowing to process large amounts of data, these methods inevitably introduce errors and the quantitative results might be imprecise. Nevertheless, we believe the overall conclusions of the study, supported by both quantitative and qualitative analyses, should hold.

References

- Robert Agnew. 1992. [Foundation for a General Strain of Crime and Delinquency](#). *Criminology (Beverly Hills)*, 30(1):47–88.
- Robert Agnew. 2001. [Building on the Foundation of General Strain Theory: Specifying the Types of Strain Most Likely to Lead to Crime and Delinquency](#). *The Journal of Research in Crime and Delinquency*, 38(4):319–361.
- Alberto Alesina, Stefanie Stantcheva, and Edoardo Teso. 2018. [Intergenerational Mobility and Preferences for Redistribution](#). *American Economic Review*, 108(2):521–54.
- Gordon W Allport. 1954. *The Nature of Prejudice*. Basic Books.
- Richard J Arneson. 1997. Egalitarianism and the Undeserving Poor. *The Journal of Political Philosophy*, 5(4):327–350.
- Mana Ashida and Mamoru Komachi. 2022. [Towards automatic generation of messages countering online hate speech and microaggressions](#). In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 11–23, Seattle, Washington (Hybrid). Association for Computational Linguistics.
- Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. *Findings of the Association for Computational Linguistics: EMNLP 2020*.
- Barcelona City Council. 2005. Ordenança de mesures per fomentar i garantir la convivència ciutadana a l’espai públic de Barcelona Acord del Consell Plenari de 23-12-2005. Article 58.2a.
- Gary S. Becker. 1968. [Crime and Punishment: An Economic Approach](#). *The Journal of Political Economy*, 76(2):169–217.
- Tom Bourgeade, Alessandra Teresa Cignarella, Simona Frenda, Mario Laurent, Wolfgang Schmeisser-Nieto, Farah Benamara, Cristina Bosco, Véronique Moriceau, Viviana Patti, and Mariona Taulé. 2023. A multilingual dataset of racial stereotypes in social media conversational threads. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 686–696.
- Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 160–172. Springer.
- Youngjin Chae and Thomas Davidson. 2023. Large language models for text classification: From zero-shot learning to fine-tuning. *Open Science Foundation*.
- Tessa ES Charlesworth, Victor Yang, Thomas C Mann, Benedek Kurdi, and Mahzarin R Banaji. 2021. Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science*, 32(2):218–240.
- Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. 2020. [An annotated corpus for sexism detection in French tweets](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1397–1403, Marseille, France. European Language Resources Association.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.
- Gerber Claudia, Alexander D Klemm, Li Liu, and Victor Mylonas. 2018. [Personal Income Tax Progressivity: Trends and Implications](#). IMF.
- Adela Cortina. 2022. *Aporophobia. Why we reject the poor instead of helping them*. Princeton University Press.
- M. Costantini, I. Meco, and A. Paradiso. 2018. Do Inequality, Unemployment and Deterrence Affect Crime over the Long Run? *Regional Studies*, 52(4):558–571.
- John Creamer, Emily Shrider, Kalee Burns, and Frances Chen. 2022. Poverty in the US: 2022. Technical report, United States Census Bureau.
- G. Curto, M.F Jojoa Acosta, F. Comim, and B. Garcia-Zapirain. 2022. Are the poor being discriminated against on the Internet? A machine learning analysis using Word2vec and GloVe embeddings to identify aporophobia. *AI & Society*.
- Thomas Davidson, Dana Warmesley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In

- Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.
- Matthew Desmond. 2023. *Poverty, by America*. Crown.
- John Hart Ely. 1980. *Democracy and distrust : a theory of judicial review*. Harvard University Press.
- Eurostat. 2022. [Database - Income and living conditions](#).
- David Everatt. 2009. The Undeserving Poor: Poverty and the Politics of Service Delivery in the Poorest Nodes of South Africa. *Politikon*, 35(3):293–319.
- Laura Faragó, Dávid Ferenczy-Nyúl, Anna Kende, Péter Krekó, and Zoltán Gurály. 2022. Criminalization as a justification for violence against the homeless in Hungary. *The Journal of Social Psychology*, 162(2):216–230.
- Antske Fokkens, Nel Ruigrok, Camiel Beukeboom, Gagestein Sarah, and Wouter Van Atteveldt. 2018. Studying Muslim stereotyping through microportrait extraction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.
- Kathleen C Fraser, Svetlana Kiritchenko, and Isar Nejadgholi. 2022. Computational modeling of stereotype content in text. *Frontiers in Artificial Intelligence*, 5:826207.
- R.B. Freeman. 1999. The Economics of Crime. *Handbook of Labor Economics*, 3:3529–3571.
- Sarah Geraghty. 2015. How the criminalization of poverty has become normalized in American culture and why you should care. *Mich. J. Race & L.*, 21:195.
- Anthony G Greenwald, Brian A Nosek, and Mahzarin R Banaji. 2003. Understanding and Using the Implicit Association Test. *Journal of Personality and Social Psychology*, 85(2):197–216.
- Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.
- E. Hinton. 2017. *From the war on poverty to the war on crime: the making of mass incarceration in America*. Harvard University Press, Cambridge, Massachusetts.
- Othman Istaiteh, Razan Al-Omouh, and Sara Tedmori. 2020. [Racist and sexist hate speech detection: Literature review](#). In *2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*, pages 95–99.
- Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.
- Svetlana Kiritchenko, Georgina Curto Rex, Isar Nejadgholi, and Kathleen C. Fraser. 2023. [Aporophobia: An overlooked type of toxic language targeting the poor](#). In *Proceedings of the 7th Workshop on Online Abuse and Harms (WOAH)*, pages 113–125, Toronto, Canada. Association for Computational Linguistics.
- Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C Fraser. 2021. Confronting abusive language online: A survey from the ethical and human rights perspective. *Journal of Artificial Intelligence Research*, 71:431–478.
- Adam Looney and Nicholas Turner. 2018. Work and opportunity before and after incarceration. Technical report, Brookings.
- Yousri Marzouki, Eliza Barach, Vidhushini Srinivasan, Samira Shaikh, and Laurie Beth Feldman. 2020. The dynamics of negative stereotypes as revealed by tweeting behavior in the aftermath of the charlie hebdo terrorist attack. *Heliyon*, 6(8):e04311.
- R.J. 2013 Miller. 2013. "Race, Hyper-Incarceration, and US Poverty Policy in Historic Perspective". *Sociology Compass*, 7, no. 7:573.
- Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J.Niels Rosenquist. 2011. Understanding the Demographics of Twitter Users. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*.
- Michael Mueller-Smith. 2015. The Criminal and Labor Market Impacts of Incarceration. *Working Paper*.
- Heather Nunn and Anita Biressi. 2009. The undeserving poor. *Soundings*, 41.
- D. Pager. 2003. The mark of a criminal record. *American Journal of Sociology*, 108(5):937–975.
- Carla Perez Almendros, Luis Espinosa Anke, and Steven Schockaert. 2020. [Don't patronize me! an annotated dataset with patronizing and condescending language towards vulnerable communities](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5891–5902, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. 2021. "call me sexist, but...": Revisiting sexism detection using psychological scales and adversarial samples. In *Proceedings of the international AAAI conference on web and social media*, volume 15, pages 573–584.
- Michael J. Sandel. 2020. *The tyranny of merit*. Penguin Random House.
- Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

- Xi Song, Catherine G. Massey, Karen A. Rolf, Joseph P. Ferrie, Jonathan L. Rothbaum, and Yu Xie. 2020. [Long-term decline in intergenerational mobility in the United States since the 1850s](#). In *Proceedings of the National Academy of Sciences of the United States of America*, volume 117, pages 251–258. National Academy of Sciences.
- Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. [Generating counter narratives against online hate speech: Data and strategies](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online. Association for Computational Linguistics.
- Juan Terradillos Basoco. 2020. [Aporofobia y plutofilia : la deriva jánica de la política criminal contemporánea](#).
- The World Bank. 2022. [Poverty and Shared Prosperity 2022: Correcting Course](#). Technical report.
- The World Bank. 2023. [Poverty and Inequality Platform \(PIP\)](#).
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *arXiv preprint arXiv:2307.09288*.
- United Nations. 2018. [Report of the Special Rapporteur on extreme poverty and human rights on his mission to the United States of America](#).
- Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. [Challenges and frontiers in abusive content detection](#). In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.
- Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini. 2020. [The role of artificial intelligence in achieving the Sustainable Development Goals](#). *Nature Communications* 2020 11:1, 11(1):1–10.
- Anton Voronov, Lena Wolf, and Max Ryabinin. 2024. [Mind your format: Towards consistent evaluation of in-context learning improvements](#). *arXiv preprint arXiv:2401.06766*.
- Zhiqiang Wang, Yiran Pang, and Yanbin Lin. 2023. [Large language models are zero-shot text classifiers](#). *arXiv preprint arXiv:2312.01044*.
- Zeeraq Waseem. 2016. [Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter](#). In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142.
- D. Wu and Z. Wu. 2012. [Crime, Inequality and Unemployment in England and Wales](#). *Applied Economics*, 44(29):3765–3775.
- Wanzheng Zhu and Suma Bhat. 2021. [Generate, prune, select: A pipeline for counterspeech generation against online hate speech](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 134–149, Online. Association for Computational Linguistics.

A Prompting LLMs

Here, we list the best zero-shot and few-shot prompts for the three LLMs.

GPT-3.5-turbo

Zero-shot prompt: Your task is to determine whether a given text associates poor people and criminality. Answer ‘yes’ if the text implies that poor or homeless people commit crimes, are arrested, go to jail, or are harassed by police. Answer ‘no’ if the text implies that poor or homeless people are the victims of crimes, if people steal from the poor, or if the text is neutral. Does the following text relate poverty, criminality, and law enforcement? Answer yes or no only. TEXT: <sentence>

Few-shot prompt: Your task is to determine whether a given text associates poor people and criminality. Answer ‘yes’ if the text implies that poor or homeless people commit crimes, are arrested, go to jail, or are harassed by police. Answer ‘no’ if the text implies that poor or homeless people are the victims of crimes, if people steal from the poor, or if the text is neutral.

TEXT: Poor people are all criminals.
ANSWER: Yes

TEXT: Jails are only for poor people.
ANSWER: Yes

TEXT: Police arrest only poor people.
ANSWER: Yes

TEXT: Police harass homeless people and destroy homeless encampments.
ANSWER: Yes

TEXT: High crime rates and high homelessness are two main issues in our country.

ANSWER: No

TEXT: Stealing from the poor is unacceptable.

ANSWER: No

Does the following text relate poverty, criminality, and law enforcement? Answer yes or no only.

TEXT: <sentence>

ANSWER:

Llama 2

System Prompt: Your task is to determine whether a given text implies that poor people commit more crimes, or should be punished for being poor. You may only output 'yes' or 'no'

Zero-shot prompt: Does the following text relate poverty, criminality, and law enforcement? Answer yes or no only. TEXT: <sentence>

Few-shot prompt: Your task is to determine whether a given text associates poor people and criminality. Answer 'yes' if the text implies that poor or homeless people commit more crimes, are arrested more, go to jail more, or are harassed by police just because they are poor. Answer 'no' if the text implies that poor or homeless people are the victims of crimes, if people steal from the poor, or if the text is neutral. Here are some examples for you to learn from:

TEXT: Poor people are all criminals.

LABEL: Yes

TEXT: Jails are only for poor people.

LABEL: Yes

TEXT: Stealing from the poor is unacceptable.

LABEL: No

TEXT: Police arrest only poor people.

LABEL: Yes

TEXT: Police harass homeless people and destroy homeless encampments.

LABEL: Yes

TEXT: High crime rates and high homelessness are two main issues in our country.

LABEL: No

Based on these examples, does the following text relate poverty, criminality, and law enforcement? Answer yes or no only.

TEXT: <sentence>

ANSWER:

Flan-T5

Zero-shot prompt: Your task is to determine whether a given text associates poor people and criminality. Answer 'yes' if the text implies that poor or homeless people commit more crimes, are arrested more, go to jail more, or are harassed by police just because they are poor. Answer 'no' if the text implies that poor or homeless people are the victims of crimes, if people steal from the poor, or if the text is neutral. Answer 'yes' or 'no' for the following text: <sentence>

Few-shot prompt: Your task is to determine whether a given text implies that poor people commit more crimes, or should be punished for being poor.

TEXT: Poor people are criminals.

ANSWER: Yes

TEXT: Jails are for poor people.

ANSWER: Yes

TEXT: Stop stealing from the poor.

ANSWER: No

TEXT: Police arrest only poor people.

ANSWER: Yes

TEXT: Police harassed homeless people, destroyed homeless encampments.

ANSWER: Yes

TEXT: Poor people are being harassed,

beaten, arrested by the police.

ANSWER: Yes

TEXT: It's scary to walk past homeless people because they will probably attack me.

ANSWER: Yes

Based on these examples, does the following text relate poverty, criminality, and law enforcement? Answer 'yes' if the text implies that poor or homeless people commit more crimes, are arrested more, go to jail more, or are harassed by police just because they are poor. Otherwise, answer 'no'.

TEXT: <sentence>

ANSWER:

Discovering Implicit Meanings of Cultural Motifs from Text

Anurag Acharya^{1,2} Diego Casto Estrada¹ Shreeja Dahal¹
W. Victor H. Yarlott¹ Diana Gomez¹ Mark A. Finlayson¹

¹ Florida International University, Miami, FL
markaf@fiu.edu

² Pacific Northwest National Laboratory, Richland, WA
anurag.acharya@pnnl.gov

Abstract

Motifs are distinctive, recurring, widely used idiom-like words or phrases, often originating in folklore and usually strongly anchored to a particular cultural or national group. Motifs are significant communicative devices across a wide range of media—including news, literature, and propaganda—because they can concisely imply a large set of culturally relevant associations. One difficulty of understanding motifs is that their meaning is usually implicit, so for an out-group person the meaning is inaccessible. We present the Motif Implicit Meaning Extractor (MIME), a proof-of-concept system designed to automatically identify a motif’s implicit meaning, as evidenced by textual uses of the motif across a large set data. MIME uses several sources (including motif indices, Wikipedia pages on the motifs, explicit explanations of motifs from in-group informants, and news/social media posts where the motif is used) and can generate a structured report of information about a motif understandable to an out-group person. In addition to a variety of examples and information drawn from structured sources, the report includes implicit information about a motif such as the type of reference (e.g., a person, an organization, etc.), its general connotation (strongly negative, slightly negative, neutral, etc.), and its associations (typically adjectives). We describe how MIME works and demonstrate its operation on a small set of manually curated motifs. We perform a qualitative evaluation of the output, and assess the difficulty of the problem, showing that explicit motif information provided by cultural informants is critical to high quality output, although mining motif usages in news and social media provides useful additional depth. A system such as MIME, appropriately scaled up, would potentially be quite useful to an out-group person trying to understand in-group usages of motifs, and has wide potential applications in domains such as literary criticism, cultural heritage, marketed and branding, and intelligence analysis.

1 Introduction

Motifs can be simply described as recurring cultural “memes” that are grounded in stories. Motifs often originate in folklore, but are ubiquitous and can be found anywhere that language is influenced by culture. They are interesting and useful because they provide a compact source of cultural information: they concisely communicate a constellation of related cultural ideas, associations, assumptions, and knowledge. One common western motif that illustrates the importance and information density of motifs is *troll under the bridge*. One folktale containing the motif, *The Three Billy Goats Gruff*, involves a troll, hiding under a bridge, who tries to devour the goats as they attempt to cross. The motif is found across the folklore of Northern Europe, especially Norway. To members of many western cultures, invoking this motif brings a number of related ideas to mind that are not directly communicated by the surface meaning of the words: the bridge is along the critical path of the heroes and they must cross it to achieve their goal; the troll lives under the bridge, surprising those who attempt to cross it; the troll tries to kill, eat, or otherwise extract some value from the would-be crossers; the troll is a squatter, not the officially sanctioned master of the bridge; and the troll usually meets his end at the hands the hero. The modern utility of the motif as a communicative device is clearly visible in the common term *patent troll*, a person or organization that claims illegitimate ownership over ideas and attempts to extract value from companies who have related products related to those ideas. Here we see an analogical transfer of cultural attributes from the troll of folklore to the “troll” of patents.

Although the example above is drawn from folklore, motifs have importance beyond folktales: they occur in modern stories, news articles, opinion pieces, press releases, propaganda, novels, movies, plays—indeed, anywhere that culture impinges on

language. One powerful modern example is the use of the *Pharaoh* motif in modern Middle Eastern discourse. The Pharaoh, which refers to the Pharaoh who opposes Moses in the narrative found in Qu’ran, is an arrogant and obstinate tyrant who oppresses the chosen people, defies the will of God, and is punished for it. In modern Islamist extremist discourse, the term *Pharaoh* has been invoked against leaders such as Anwar Sadat of Egypt, Ariel Sharon of Israel, and George W. Bush, the last of whom Osama bin Laden referred to as the “pharaoh of the century” (Halverson et al., 2011). Without understanding the implications of the *Pharaoh* motif, we would be unable to understand both the content of this message (that these leaders are being cast as oppressors) and the cultural group for whom this message was intended (the chosen people, in this case the Ummah).

Work in cognitive psychology has shown that in-group members understand the associations for motifs from their group better than out-group persons (Acharya, 2022). Critically, people can recognize that a motif is being used, regardless of their group membership, and out-group speakers recognize that those motifs carry special meaning and are aware that they’re missing out on that meaning. A natural next step, then, is some way of making motif meanings available to out-group persons.

Unfortunately, current resources describing motifs are not sufficient for out-group people to understand the deeper meaning and context of a motif. Let us take an example of a Hindu motif, *Saraswati*. If one looks up *Saraswati* on Google or Wikipedia, one can read that she is the goddess of knowledge, wisdom, music and art, among other things. But this gives no insight into the meaning of the common phrase “Kartik has Saraswati on his tongue.” It might sound like that means Kartik is highly knowledgeable or wise, but what it actually means is Kartik never tells a lie. Existing resources, like motif indices, dictionary definitions, and encyclopedia entries do not usually bring forth the nuance and the gravity of meaning that motifs typically carry. A number of additional sources of information are needed, including example usages, to provide the larger context of motif meaning.

To address this problem, we propose the *Motif Implicit Meaning Extractor* (MIME), a proof-of-concept system that can extract and organize various implicit meaning of motifs for presentation to an out-group audience for better understanding.

The MIME takes as a query a chunk of text that contains a motif of interest, and uses information from various sources like encyclopedia entries, explanations by in-group informants, and online discourse (e.g., news or social media) to generate a structured report on the common meaning and usage of the motif. Appendix A shows a sample MIME report.

The paper is organized as follows. We first review related work on motifs and their computational processing (§2). Then we list the kinds of implicit information that we would like to extract (§3, followed by the data we collected to support this extraction (§4). Next we explain in detail the system architecture (§5). We then describe our qualitative evaluation, showing that critical information presented in the reports is more likely present in explicit in-group explanations and implicit usage, rather than in Wikipedia entries or news usages (§5.5). Finally, we discuss limitations of this proof-of-concept and map out future work (§7).

2 Related Work

2.1 Motifs

Stith Thompson informally defined a motif as an item “worthy of note because of something out of the ordinary, something of sufficiently striking character to become a part of tradition, oral or literary. Commonplace experiences, such as eating and sleeping, are not traditional in this sense. But they may become so by having attached to them something remarkable or worthy of remembering.” (Thompson, 1960, p. 19). In folklore, motifs are preferentially retained throughout retellings and recombinations of tales due to their striking nature and the density of information they communicate. Folklorists have long hypothesized that a tale’s specific composition of motifs can be used to trace the tale’s lineage (Thompson, 1977, Part 4, Chapter V). This has led folklorists to construct motif indices that identify motifs and note their presence in specific tales (usually as represented in a particular folkloristic collection). The most well-known motif index is the Thompson motif index (Thompson, 1960). Thompson’s index designates each motif with a code; for example, *troll under a bridge* is referenced by the codes G304 and G475.2. In this case, *troll under a bridge* is represented by two motifs as Thompson generalizes trolls to ogres, a general class of monstrous beings; thus, the motifs are *troll as ogre* (G304) and *ogre attacks intruders on bridge* (G475.2). Thompson

noted that motifs generally fall into one of three subcategories (Thompson, 1977, pp. 415–416): events, characters, or props. Examples of these include *Hero rescuing a Princess* (B11.11.4; event), *Old Man Coyote* (A177.1, character), and *Magic Carpet* (D1155, prop).

Folklorists were the originators of the idea of motif, and constructed motif indices that identify motifs and note their presence in specific folktales. As stated previously, the most well-known motif index is Thompson’s motif index (TMI) (Thompson, 1960), which integrates a number of prior indices and references folktales from over 600 collections, indexed to 46,248 motifs and submotifs. In addition to this, Thompson provides substantial discussion on motifs and the compilation of motif indices in his book *The Folktale* (Thompson, 1977). Additionally, there are many motif indices which target specific cultures and periods, for example, early Irish literature (Cross, 1952), traditional Polynesian narratives (Kirtley, 1971), Japanese folk literature (Ikeda, 1971), or early Icelandic literature (Boberg, 1966). While Thompson’s motif index is perhaps the primary source of motif information used today, it has been criticized because of overlapping motif subcategories, censorship (primarily of obscenity), and missing motifs (Dundes, 1997).

2.2 Computational Approaches to Motifs

Darányi (2010) called attention to the need for research into the automation of extraction and annotation of motifs in folklore, and suggested that motifs have application in storing, indexing, and retrieving documents based on the motifs contained within. Work has also been done examining the shortcomings and potential applications of motifs. For example, Darányi and Forró (2012) determined, based on cluster analysis, that motifs may not be the highest level of abstraction in narrative, echoing criticisms that many motifs are interdependent (Dundes, 1997). Darányi et al. (2012) have made substantial headway towards using motifs as sequences of “narrative DNA”, and Ofek et al. (2013) have demonstrated learning tale types based on these sequences. Declerck et al. (2012) have also done work on converting electronic representations of TMI and ATU (Uther, 2004) to a format that enables multilingual, content-level indexing of folktale texts, building upon past work (Declerck and Lendvai, 2011). Currently, this work appears to

be focused on the descriptions of motifs and tale types, without reference to the stories.

2.3 Relation Detection, Information Extraction, and Template Filling

The MIME can be thought of as a mix of targeted information extraction, relationship extraction, and template filling, all tasks that have seen much attention in the past.

Soares et al. (2019) used distributional similarity to build a general relation extraction system using BERT (Devlin et al., 2018). Similarly, Wu and He (2019) also used BERT to perform relation classification while also incorporating information about large entities. Meanwhile Ye and Ling (2019) and Wang et al. (2016) implement CNN-based few-shot relation classifiers. There are several other systems that perform well on relation detection, such as Bastos et al. (2021), a neural network based model; Kim et al. (2019), an RNN based model; and Cai et al. (2016), a BRCNN based model.

There has also been much work on information extraction (IE), both in general and specific types of relations. One important work is OpenIE by Stanovsky et al. (2018), a system which re-frames open information extraction as a sequence tagging problem. Similar systems include those by Cetto et al. (2018), Gashteovski et al. (2017), and Bhutani et al. (2016) which all perform close to the state of the art. Importantly, OpenIE defined a standard relation set, and certain systems have focused on improving performance on specific relations in that set. For example, Pal et al. (2016) focuses on nominal OpenIE, which finds an efficient way to extract open relations for compound noun phrases. Similarly, Saha et al. (2017) focuses specifically on numerical relations to extract OpenIE tuples, and Saha et al. (2018) addresses the issue of extracting relation tuples for conjunctive sentences.

It is worth noting that we did try to see if our task could be achieved by using an existing off-the-shelf IE systems. We tested several state-of-the-art IE systems including OpenIE (Stanovsky et al., 2018), CALMIE, (Saha et al., 2018) and BONIE (Saha et al., 2017). We found the output of these systems were not adequate to the task, which we traced to the fact that most IE systems focus on each sentence in isolation, i.e., without context. Taking advantage of context required breaking out individual IE subtasks in a different way, which led to the MIME approach described in Section 5.

Finally, a number template filling systems are related to the MIME task. For example, [Jean-Louis et al. \(2011\)](#) combines text segmentation and graph techniques to perform template filling. Another is [Miliani et al. \(2019\)](#) which splits text into frames in order to accomplish slot filling. But perhaps the work that most closely relates to this work is work done by [Chambers and Jurafsky \(2011\)](#) which combines the tasks of information extraction and template filling, but is able to do so without having a fixed template for the output in advance. For this work, we draw on all these tasks, mixing and matching our implementations to maximize explainability of the proof-of-concept system. In particular, we defer the use of black-box neural systems, which can be used to optimize performance, for later implementations.

3 Types of Implicit Information

There are several categories of implicit information, listed below, we would like to expose about motifs to improve understanding by out-group persons.

3.1 Type of Reference

The first piece of information that is useful to understand a motific usage is the answer to the question “What kind of object does this motif refer to?” Thompson roughly split motifs into three broad types: events, characters, or props. Which one a motif refers to may not always be obvious from context. For example, “That person is such an *Amalek*.” makes clear that the object being referred to is a person, but the usage “Sorry, Jack: They are bad, but they are no *Amalek*.” or even “What an *Amalek*.” is ambiguous.

3.2 General Connotation

The second piece of implicit information that is useful is the general positive, neutral, or negative connotation of the motif, which may unknown or be different than commonly assumed. In the “That person is such an *Amalek*.” example above, the connotation alone can give us significant insight into the meaning. *Amalek* refers to an ancient ruler in the Hebrew Bible who was a well-known persecutor of the Jews, and is usually used to speak negatively about a person. Another good example is the motif *Leprechaun*: while in the broader Western culture it is assumed to have a general positive connotation, within an Irish context specifically it is more neutral, being a mix of positive and neg-

ative elements, as Leprechauns are mischievous, tricky beings, but not necessarily evil.

3.3 Associations

Finally, a deeper understanding of a motif involves understanding the specific associations or implications it calls to mind. With the example above of *troll under the bridge* in the introduction, there are a variety of implications about the motives and legitimacy of the troll that the motif calls to mind. As another example, in the case of *Finn McCool*, using this motif to describe a person implies potentially that they are incredibly smart or powerful, or act as a savior of many people or protector of the land.

4 Data

4.1 Motif Selection

There are literally tens of thousands of motifs listed within the many hundreds of motif indices that have been written since the late 1800s. Most of these motifs, furthermore, are not actually well known or commonly used within the relevant group. This sparsity presents a challenge for identifying motifs that are actually used in communicative context of interest. It was not our goal in this work to solve this particular problem. Instead, our goal was to develop a method that, given a list of motifs known to be in use, could extract the relevant implicit information. Therefore, we first had to manually select a set of target motifs with which we could develop our proof-of-concept system.

We determined three criteria for selecting motifs on which to demonstrate the MIME. First, they needed to be “high quality” motifs from a clearly identifiable group, which we define below. Second, we needed access to members of the relevant group to gather in-group explanations of motif meaning, to be used both as system data and evaluation materials. Finally, the number of motifs needed to be small enough to be manageable within the project scope and budget. We defined “high quality” motifs according to the following three characteristics:

1. **Common Use:** Selected motifs should be in common use in modern communicative contexts. The simplest test of this was a keyword search to see if the actual words expressing the motif were used either on social media, such as Twitter, or in the news. This criteria simplified the search—if we couldn’t find it, we did not include it.

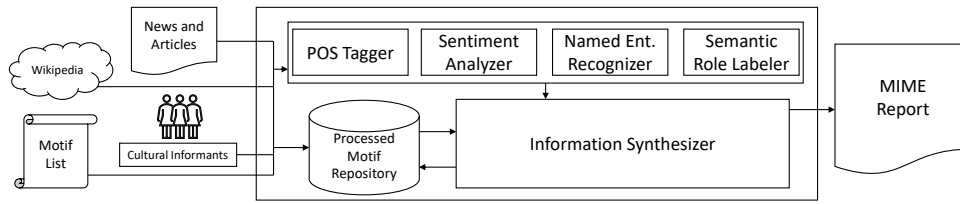


Figure 1: The main architecture design of the Motif Implicit Meaning Extractor (MIME), including the external components that feed into the system.

2. **Clear Source:** Selected motifs must have a clearly identifiable source within the group. By *source*, we mean an associated, well-known story within the same body of folklore as other motifs for the group. This criteria is intended to provide proof of relevance for the motif to the group in question. If a motif had no definitive source within the folklore of the cultural group we excluded it.
3. **Strength:** Selected Motifs should be preferentially used in a motific way, meaning their usage usually draws on the implicit meaning, rather than being used literally or eponymically (i.e., as a name). This was a subjective judgement of *how* the motifs were used when we found them in news.

Intersecting these three criteria led us to select roughly ten motifs from each of three groups: Irish, Jewish, and Puerto Rican. These were groups that had easily accessible motif indices, quite of a bit of activity by group members on news and social media, and whose members were accessible to us through various connections and contacts. Regarding the motif indices, for Irish, we used T.P. Cross’s *Motif-Index of Early Irish Literature* (Cross, 1952) as a main source; for Puerto Rican, we drew motifs from S.R. Lamarche’s *The Mythology and Religion of the Tainos* (Hurley et al., 2021), R.E. Alegría’s *The Three Wishes: A Collection of Puerto Rican Folktales* (Alegría et al., 1969), and J. Ramírez-Rivera’s *Puerto Rican Tales: Legends of Spanish Colonial Times* (Ramírez-Rivera et al., 1977); and for Jewish motifs, we referenced D.N. Noy’s *Motif-index of Talmudic-Midrashic literature* (Noy, 1954). We read through these indices, and in discussion with our in-group informants and investigation on social media and news search engines, we settled on the following set:

Irish (13) The Salmon of Wisdom, Finn McCool, Leprechaun, King Conchobar, Aos Si, Ban-

shee, Cu Chulainn, the Wren, the Magic Harp, Tir Na Nog, Shamrock, Fairy Fort, the Children of Lir

Jewish (9) Haman, Golem, Amalek, Babel, Leviathan/Behemoth, 70 Languages, Name in Vain, the Ark of the Covenant, Kiddush Hashem

Puerto Rican (12) Reyes Magos/Three Kings, Agueybana, Atabey, Roberto Cofresi, Divina Providencia, Guanina, Juan Bobo, Yocahu, the Coqui, Hormigueros, Jibaro/Jibarito, Chupacabra

4.2 Wikipedia Articles

Many motifs have corresponding Wikipedia articles which contain some useful information (although, as noted above, are not by themselves sufficient to understand motif usages). For each motif in our set we found and downloaded the relevant Wikipedia article if it existed (19 out of 34). We converted the article to plain text, eliminating formatting and surrounding elements (e.g., links to other pages, donation advertisements). We further eliminated irrelevant article sections of the articles (e.g., References, History), while retaining sections like Background Information, Summary, and Traditions. Although we only looked at Wikipedia in this work, it would be possible to expand this portion of the data by adding other encyclopedias, thesauri, or dictionaries.

4.3 In-Group Informant Explanations

For each motif we also interviewed our in-group informants as to their meaning. In these short interviews we provided some example usages of the motif (found via keyword search) and asked our informants to explain the meaning of the motif in context. We produced short descriptions that summarized the answers of multiple informants, which were then verified by the informants themselves. Examples for three motifs are as follows:

Finn McCool (Irish, Character) Incredibly smart, powerful savior of many people. Considered a protector of the land.

Kiddush (Jewish, Event) A prayer; more commonly invoked as “kiddush hashem” which is representing God, the Torah, and the Jewish people in the best light possible.

Coqui (Puerto Rican, Prop) Small frog species native to Puerto Rico. Known for being the national symbol of Puerto Rico.

4.4 Online News

We also collected a large number of online news articles where the motifs were used in a motific sense. We obtained English texts through NexisUni¹, a university version of LexisNexis, a tool for searching through news articles, which provides worldwide scope for news and related text. We searched for motif terms and batch downloaded these articles, as allowed by the University’s license. These articles were then further processed by a Lucene-based lexical matcher, with fuzzy rules for a variety of lexical forms for each motif, to verify the presence of motif terms (called *motif candidates*) for inspection by our annotators. This dataset comprises 26,078 motifs candidates across 7,955 texts.

We hired annotators who identified as members of the groups in question, which was determined through an interview. We also required annotators to possess a college degree and be fluent in English. We hired six annotators total (two annotators per group) to perform the double-blind annotations. Annotators were asked, for each motif candidate, to identify whether or not the usage invoked the implicit meaning of a motif (e.g., referring to something large and monstrous as a *behemoth*). A motif candidate that invoked the in-group meaning was called a *motif instance*. The Jewish and Puerto Rican teams participated produced motif instance annotations with an average agreement of $\kappa > 0.7$ while the Irish team produced annotations with an average agreement of $\kappa > 0.55$. The annotation resulted in 1,723 motif instances (159 Irish, 1,215 Jewish, and 349 Puerto Rican).

5 Computational Pipeline

We implemented a traditional NLP pipeline, rather than an end-to-end neural system (generative or

¹<https://www.lexisnexis.com/en-us/professional/academic/nexis-uni.page>

otherwise) for several reasons. First, a pipelined system is more explainable, allow us to do error analysis as to whether individual pieces of information were easier or harder to extract, allowing us to refine our task definition. Second, we have relatively little data with which to fine-tune a neural system. In particular, due to the difficulty of collecting high-quality information about specific motifs, (a) we have only 34 motifs in our demonstration, (b) we have only 1,723 annotated motif instances (159 Irish, 1,215 Jewish, and 349 Puerto Rican), and (c) the in-group informant explanations are only a few sentences long each. Third, since we are concentrating on the structure of the task and not optimizing for performance, a pipelined system is more useful; once the task is well defined and more data has been collected, later work can focus on integrating neural architectures for optimization. Finally, our anecdotal experiments with the latest generative text models (e.g., chatGPT) found that it was quite poor in its ability to generate accurate information about motifs.

The proof-of-concept MIME system has four components, described below. We applied all components to all of the data, including the Wikipedia articles, in-group informant explanations, and 100-token windows surrounding annotated motif instances in the online news. Information extracted from these data are stored in a database indexed to each motif, and can be used to generate a motif report (as shown in Appendix A).

5.1 Part-of-Speech Tagger

Part-of-Speech (POS) tagging is an important piece of information for understanding type of reference and associations. We used SpaCy’s (Honnibal et al., 2020) medium-sized CPU-optimized POS tagger.

5.2 Named Entity Recognizer

The types of Named Entities (NEs) mentioned in relation to a motif again speaks to type of reference and associations. We used an implementation of the baseline Named Entity Recognizer (NER) model from (Peters et al., 2017), which was provided by AllenNLP in the form of its ELMo-Based NER. This model achieves a reported score of 96% F1 score on the CoNLL-2003 validation set (Tjong Kim Sang and De Meulder, 2003). The AllenNLP NER uses twenty-one classes, organized into a hierarchy. We reduce the number of possible NER tags for the motifs by consolidating named entities tags into their top-level parent class. This resulted in

just four classes: Person, Organization, Location, and Miscellaneous.

5.3 Sentiment Analyzer

Sentiment is critical to understanding the implicit connotation of a motif. We used AllenNLP’s sentiment analyzer, which outputs either a 0 or a 1 using an LSTM classifier with GloVe embeddings (Pennington et al., 2014) with a reported performance of 87% accuracy on the Stanford Sentiment Treebank corpus (Socher et al., 2013), which is near to state-of-the-art performance.

5.4 Semantic Role Labeler

Semantic role labeling exposes subject-relation-object structures within the texts, which is relevant to the type of reference and implicit associations. We used AllenNLP’s Semantic Role Labeler (SRL), which is a slightly modified version of a BERT based model (Shi and Lin, 2019), with a reported F_1 of 86.49 on the Ontonotes 5.0 (Weischedel et al., 2013). While AllenNLP’s SRL generates role labels in the full PropBank scheme (Babko-Malaya, 2005), this level of complexity was not necessary for our proof-of-concept system. Therefore we mapped the various SRL tags into just three main roles: ARG0 (agent), ARG1 (patient), and ARG2 (location/instrument/beneficiary/etc.), while ignoring all other roles as “non-specific”.

5.5 Information Synthesizer

Information generated by the prior four components over all the data is integrated together in the information synthesizer modules. First, NER taggings are examined from across the all data to determine the most common type reference, which falls into one of the four categories of *Person*, *Organization*, *Location*, and *Miscellaneous*.

Second, the sentiment values across all the data for each motif are averaged and transformed into one of four broad categories, as shown in Table 1.

Sentiment Label	Value
<i>Negative</i>	0 – 0.25
<i>Slightly Negative to Neutral</i>	0.25 – 0.5
<i>Neutral to Slightly Positive</i>	0.5 – 0.75
<i>Positive</i>	0.75 – 1

Table 1: New sentiment categories for the motifs

Third, the SRL Subject-Verb-Object triples for each motif are processed. When the subject or object contains the motif of interest, we look at the

1. Motif Type: Whether a motif is event, character, or prop
2. Origin Culture: What culture the motif originated from
3. Usually referred to: What type of entity the motif refers to, and how it’s used in a sentence
4. Major Associations: What are the major associations of the motif
5. General Usage Connotation: How is the motif typically perceived?
6. Examples: Example usages of the motif
7. Background: Short explanation of the motif

Table 2: Templates to be filled out for the MIME report

other role, and if this is an adjective, this is listed as an association of that motif. If the other role does not contain an adjective, we return to the original sentence to see if there are any adjectives in the sentence within a 21-word window centered on the motif, and include these as associations.

With all the above information processed from the data and included in the MIME’s database, the MIME can produce a report for each motif of interest. We treated report generation as a specialized template-filling task. The template fields are shown in Table 2. To match the fields in the template, all the information in the synthesizer goes through the process of converting tags and stubs into phrases and/or sentences and merging information together from different modules to put relevant information together. One possible source of error in this process is if the information from the external sources or the catalog is too long. To avoid this, we truncate the information to a fixed length. The final result of this process is a filled out template, which is the final MIME Report, as shown in Appendix A.

6 Results and Discussion

Since this task is a version of targeted Information Extraction and the output is a human-readable report, it is challenging to apply traditional evaluation metrics to assess effectiveness. We performed a qualitative analysis of the reports, comparing them against the explanations provided by the in-group informants as well as our own general understanding of the meaning of motifs.

In general, MIME did well at finding implicit information for motifs that have several instances of proper usage in news. It was able to identify the type of reference of the motifs fairly accurately (only 1 answer incorrect/missing, because of insufficient data), which is useful because this information was not provided by the in-group explanations. MIME was also able to reveal the general

connotation of the motifs correctly across most of the motifs (6 incorrect). In some cases, like *Lep-rechaun*, from usages in the news data it was able to infer a connotation closer to that provided by the in-group explanations, despite the popular understood meaning being slightly different. This shows that the MIME approach adds value beyond what is provided in the explicit in-group explanations.

We also evaluated the the proportion of information derived from our three sources. We examined the three main types of implicit information—type of reference, general connotation, and associations—and counted how often that information could be traced back to a specific type of source (Wikipedia, Explanations, or Online News). The summary of these results are shown in Tables 3 and 4. The results show that the bulk of the associations (roughly 80%) come from the in-group explanations, with approximately 14% coming from news. This result shows the importance of the in-group explanations, and that the meaning of motifs, in general, is not discoverable from usage in modern news alone. While the Wikipedia text was not particularly information dense for this task, in our qualitative analysis it played a crucial role of providing additional context in terms of clear-to-read summaries and background, thereby providing some level of utility. This is especially visible for the Puerto Rican motifs, where the Wikipedia data was especially devoid of depth, and the reports are notably less rich than the other two groups. Overall, we see that the results show the potential of this approach, consistent with its proof-of-concept stage of development.

7 Limitations and Future Work

As a proof-of-concept, MIME has a number of important limitations. First, by design the system does not attempt to extract information that must be obtained via multiple steps of inference. Second, the system sometimes reports associations that are inconsistent with those reported by the cultural informants. In our qualitative evaluation this is mostly because the motif seems to have been used incorrectly by out-group persons in online news. A few of these mistakes can be attributed to an adjective that is near the motif but was grammatically attached elsewhere. A more sophisticated approach to attribute extraction could mitigate this issue, or collecting further example usages from cultural informants and assigning higher weight to

	Wiki	Explanations	News
Reference	1.4%	3.4%	95.2%
Connotation	0.7%	2.2%	97.7%
Association	2.4%	79.5%	13.7%
Overall	1.5%	28.4%	68.9%

Table 3: Proportion of implicit information that can be attributed to each source, grouped by information type.

	Wiki	Explanations	News
Irish	2.7%	30.6%	68.0%
Jewish	1.8%	22.9%	73.1%
PR	0%	31.6%	65.6%
Overall	1.5%	28.4%	68.9%

Table 4: Proportion of implicit information that can be attributed to each source, by group

those. Finally, if the data sources lack relevant information the system will naturally fall short; this was evident in particular for the Puerto Rican motif *Yocahu*, where there were very few news usages.

One key future step is to look more comprehensively at the discourse context of the motifs instances. For instance, we not only have sentiment values and linguistic properties for the sentences with motifs themselves, but also for other sentences that form part of the discourse that involves the motif. Using this discourse relation information, we will be able to more accurately predict the associations for the motifs. We could not implement such an approach because of lack of sufficient examples of motifs in narrative text; with more data of real-world narrative usage of motifs, this could potentially be implemented in the future.

The importance of the in-group explanations for implicit associations is also revealing. Looking only at usages in modern discourse would make it very hard to extract these associations; on the other hand, asking in-group persons to explain every motif is time-consuming and laborious. Here, one untapped source of information would be the original narratives themselves, from which the original meaning of the motifs are derived. Ideally, a next step would be to mine this information for the original context and meaning of the motifs.

One main limitation of MIME, of course, is that it requires manual identification and annotation of the motifs. Recent work by [Yarlott et al. \(2021\)](#) demonstrated a preliminary motif detection system which perhaps could form the foundation for a scaling up of the MIME to as-yet-unanalyzed motifs.

Acknowledgments

This work was supported in part by DARPA contract FA8650-19-C-6017. The primary author of this research, Dr. Acharya, is now at Pacific Northwest National Laboratory, which is operated by Battelle Memorial Institute for the U.S. Department of Energy under contract DE-AC05-76RLO1830, but this work was completed while he was at FIU.

References

- Anurag Acharya. 2022. *Integrating Cultural Knowledge into Artificially Intelligent Systems: Human Experiments and Computational Implementations*. Ph.D. thesis, Florida International University.
- R.E. Alegria, R.E. Alegría, L. Homar, and E. Culbert. 1969. *The Three Wishes: A Collection of Puerto Rican Folktales*. Harcourt, Brace & World.
- Olga Babko-Malaya. 2005. Propbank annotation guidelines.
- Anson Bastos, Abhishek Nadgeri, Kuldeep Singh, Isaiah Onando Mulang, Saeedeh Shekarpour, Johannes Hoffart, and Manohar Kaul. 2021. Recon: relation extraction using knowledge graph context in a graph neural network. In *Proceedings of the Web Conference 2021*, pages 1673–1685.
- Nikita Bhutani, HV Jagadish, and Dragomir Radev. 2016. Nested propositions in open information extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 55–64.
- Inger Margrethe Boberg. 1966. *Motif-index of early Icelandic literature*. Munksgaard.
- Rui Cai, Xiaodong Zhang, and Houfeng Wang. 2016. Bidirectional recurrent convolutional neural network for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 756–765.
- Matthias Cetto, Christina Niklaus, André Freitas, and Siegfried Handschuh. 2018. Graphene: Semantically-linked propositions in open information extraction. *arXiv preprint arXiv:1807.11276*.
- Nathanael Chambers and Dan Jurafsky. 2011. Template-based information extraction without the templates. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 976–986.
- Tom Peete Cross. 1952. *Motif-index of early Irish literature*. Indiana University.
- Sándor Darányi. 2010. *Examples of Formulaity in Narratives and Scientific Communication*. In *Proceedings of the First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts*, pages 29–35.
- Sándor Darányi and László Forró. 2012. *Detecting Multiple Motif Co-occurrences in the Aarne-Thompson-Uther Tale Type Catalog: A Preliminary Survey*. *Anales de Documentación*, 15(1).
- Sándor Darányi, Peter Wittek, and László Forró. 2012. Toward Sequencing “Narrative DNA”: Tale Types, Motif Strings and Memetic Pathways. In *Third Workshop on Computational Models of Narrative (CMN)*, pages 2–10, Istanbul, Turkey. European Language Resources Association (ELRA).
- Thierry Declerck and Piroska Lendvai. 2011. Linguistic and semantic representation of the thompson’s motif-index of folk-literature. In *Research and Advanced Technology for Digital Libraries*, pages 151–158. Springer.
- Thierry Declerck, Piroska Lendvai, and Sándor Darányi. 2012. *Multilingual and Semantic Extension of Folk Tale Categories*. In *Proceedings of the 2012 Digital Humanities Conference (DH 2012)*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Alan Dundes. 1997. The motif-index and the tale type index: A critique. *Journal of Folklore Research*, pages 195–202.
- Kiril Gashteovski, Rainer Gemulla, and Luciano del Corro. 2017. *MinIE: Minimizing facts in open information extraction*. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2630–2640, Copenhagen, Denmark. Association for Computational Linguistics.
- Jeffrey R Halverson, Steven R Corman, and HL Goodall Jr. 2011. *Master narratives of Islamist extremism*. Palgrave Macmillan.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. *spaCy: Industrial-strength Natural Language Processing in Python*. *Online code repository*.
- A. Hurley, C.R.R. de Arellano, and S.R. Lamarche. 2021. *The Mythology and Religion of the Tainos*. Independently Published.
- Hiroko Ikeda. 1971. *A type and motif index of Japanese folk-literature*. Orient Cultural Service.
- Ludovic Jean-Louis, Romaric Besançon, and Olivier Ferret. 2011. Text segmentation and graph-based method for template filling in information extraction. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 723–731.
- Byoungjae Kim, KyungTae Chung, Jeongpil Lee, Jungyun Seo, and Myoung-Wan Koo. 2019. A bi-lstm memory network for end-to-end goal-oriented dialog learning. *Computer Speech & Language*, 53:217–230.

- Bacil F Kirtley. 1971. *A motif-index of traditional Polynesian narratives*. University of Hawai'i Press.
- Martina Miliani, Lucia C Passaro, and Alessandro Lenci. 2019. Text frame detector: Slot filling based on domain knowledge bases. In *CLiC-it*.
- Dov Neuman Noy. 1954. *Motif-index of Talmudic-Midrashic literature*. Indiana University.
- Nir Ofek, Sándor Darányi, and Lior Rokach. 2013. [Linking Motif Sequences with Tale Types by Machine Learning](#). In *Proceedings of the 4th Workshop on Computational Models of Narrative (CMN'13)*, volume 32, pages 166–182, Hamburg, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.
- Harinder Pal et al. 2016. Demyonyms and compound relational nouns in nominal open ie. In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction*, pages 35–39.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. [Glove: Global vectors for word representation](#). In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *ACL*.
- J. Ramírez-Rivera, B. Klein, and J. Slemko. 1977. *Puerto Rican Tales: Legends of Spanish Colonial Times*. Ediciones Libero.
- Swarnadeep Saha, Harinder Pal, et al. 2017. Bootstrapping for numerical open ie. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 317–323.
- Swarnadeep Saha et al. 2018. Open information extraction from conjunctive sentences. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2288–2299.
- Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *ArXiv*, abs/1904.05255.
- Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. *arXiv preprint arXiv:1906.03158*.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.
- Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 885–895.
- Stith Thompson. 1960. *Motif-index of folk-literature: a classification of narrative elements in folktales, ballads, myths, fables, mediaeval romances, exempla, fabliaux, jest-books and local legends*, volume 4. Indiana University Press.
- Stith Thompson. 1977. *The folktale*. Univ of California Press.
- Erik F. Tjong Kim Sang and Fien De Meulder. 2003. [Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition](#). In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.
- Hans-Jörg Uther. 2004. *The types of international folktales: a classification and bibliography, based on the system of Antti Aarne and Stith Thompson*. Suomalainen Tiedeakatemia, Academia Scientiarum Fennica.
- Linlin Wang, Zhu Cao, Gerard De Melo, and Zhiyuan Liu. 2016. Relation classification via multi-level attention cnns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1298–1307.
- Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23.
- Shanchan Wu and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 2361–2364.
- W Victor H Yarlott, Armando Ochoa, Anurag Acharya, Laurel Bobrow, Diego Castro Estrada, Diana Gomez, Joan Zheng, David McDonald, Chris Miller, and Mark Alan Finlayson. 2021. [Finding trolls under bridges: Preliminary work on a motif detector](#). In *Proceedings of the Ninth Annual Conference of Advancements in Cognitive Systems (ACS) 2021*.
- Zhi-Xiu Ye and Zhen-Hua Ling. 2019. Multi-level matching and aggregation network for few-shot relation classification. *arXiv preprint arXiv:1906.06678*.

A Example Motif Implicit Meaning Extractor Report

The figure below shows a MIME report generated by the system, for the motif "*Leprechaun*". The report has been manually formatted for the sake of clarity, but the contents remain unchanged.

<i>Title</i>	Content
<i>Motif</i>	Leprechaun
<i>Found in</i>	I previously talked about the franchise in my usual breakdown format four years ago, but I've never ranked them. We live in a world where there are seven Leprechaun movies. Seven. I just can't comprehend that. They're not even particularly good movies. Of course, I say this and I own the entire series on blu-ray. Hey, they're still fun and Warwick Davis is always entertaining. So let's look at all seven of these movies and see which is the best!
<i>Motif Type</i>	Character
<i>Origin Culture</i>	Irish
<i>Usually referred to</i>	Mostly used to refer to a person. Generally used as no specific role (eg. subject/object) in a sentence.
<i>Major associations</i>	Tricky, grumpy, short, afraid
<i>General Usage Connotation</i>	Neutral to Slightly Negative
<i>Motivic Examples</i>	(1) That old miser is a real leprechaun. (2) That leprechaun at the used car lot really got the better of me. (3) I swear I could jump over Ethan, he's a real leprechaun.
<i>Referential Example</i>	A leprechaun is a type of fairy of the aos si in Irish folklore.
<i>Unrelated Example</i>	N/A—most uses are going to be at least somewhat culturally related.
<i>Eponym Example</i>	The photo you see for Leprechaun, Inc. is a 5,000 year old Dolmen, or Portal Tomb, built during the Neolithic Period.
<i>Background</i>	A leprechaun (Irish : /leipreachán/luchorpán/) is a diminutive supernatural being in Irish folklore, classed by some as a type of solitary fairy. They are usually depicted as little bearded men, wearing a coat and hat, who partake in mischief. In later times, they have been depicted as shoemakers who have a hidden pot of gold at the end of the rainbow. Leprechaun-like creatures rarely appear in Irish mythology and only became prominent in later folklore. They are usually depicted as little bearded men, wearing a coat and hat, who partake in mischief. In later times, they have been depicted as shoe-makers who have a hidden pot of gold at the end of the rainbow. Leprechaun-like creatures rarely appear in Irish mythology and only became prominent in later folklore

Can Large Language Models (or Humans) Disentangle Text?

Nicolas Audinet de Pieuchon¹ Adel Daoud^{1,2,5} Connor T. Jerzak³

Moa Johansson¹ Richard Johansson^{1,4}

¹Chalmers University of Technology, Sweden ²Linköping University, Sweden

³University of Texas at Austin, USA ⁴University of Gothenburg, Sweden

⁵The Center for Advanced Studies in Behavioral Sciences, Stanford University, USA.

{nicolas.audinet, daoud, moa.johansson, richajo}@chalmers.se,
connor.jerzak@austin.utexas.edu

Abstract

We investigate the potential of large language models (LLMs) to disentangle text variables—to remove the textual traces of an undesired forbidden variable in a task sometimes known as text distillation and closely related to the fairness in AI and causal inference literature. We employ a range of various LLM approaches in an attempt to disentangle text by identifying and removing information about a target variable while preserving other relevant signals. We show that in the strong test of removing sentiment, the statistical association between the processed text and sentiment is still detectable to machine learning classifiers post-LLM-disentanglement. Furthermore, we find that human annotators also struggle to disentangle sentiment while preserving other semantic content. This suggests there may be limited separability between concept variables in some text contexts, highlighting limitations of methods relying on text-level transformations and also raising questions about the robustness of disentanglement methods that achieve statistical independence in representation space.

1 Introduction

When computational social scientists analyze text data there are situations where the text is contaminated by a *forbidden variable* that we want to preclude from our analysis or handle in a special way. For instance, methods in causal inference that use text to correct for unseen confounders (Keith et al., 2020; Roberts et al., 2020) run into difficulties if the text is influenced by the treatment variable (Daoud et al., 2022; Gui and Veitch, 2023). In other contexts, due to ethical, legal, or robustness considerations, we may want to ensure that models trained on a corpus are not influenced by effects such as demographic factors (Bolukbasi et al., 2016), domains (Ganin and Lempitsky, 2015), personal information (Li et al., 2018), or other sensitivity information (Hovy and Prabhunoye, 2021).

There are several methods that process text representations to enforce invariance with respect to a forbidden variable (Barrett et al., 2019; He et al., 2020; Ravfogel et al., 2020; Haghhighatkhah et al., 2022; Belrose et al., 2023). With a few exceptions, these methods operate on a numerical *representation* of the text, and not directly on the text itself, which makes them less interpretable. Furthermore, they typically require a large set of annotated examples of the forbidden variable, which may not always be available or may be costly to obtain.

Computational social science using text as data is currently being transformed by the introduction of large language models (LLMs) applied in a zero-shot or few-shot fashion (Ziems et al., 2024; Törnberg, 2024). In this paper, we consider the question of whether LLMs have an out-of-the-box ability to *disentangle* a text: to transform it so that the value of a forbidden variable is hidden while preserving as much as possible of the original text. We apply the LLMs in a few-shot setup to reproduce the scenario where we have no large-scale annotations available of the forbidden variable. We investigate a variety of prompt-based techniques. If successful, the disentanglement process would also be understandable to a human reader, since the changes are carried out via interpretable transformations of the text itself.

We find that this type of disentanglement is challenging for the current generation of LLMs, as well as human annotators, to carry out. While the most powerful LLMs (e.g., GPT-4-class models) sometimes transform text so that it is difficult for humans to determine the original value of the forbidden variable, we see only a slight reduction in the accuracy of classifiers predicting the forbidden variable; its statistical presence is clearly maintained.

2 Related work

Our work builds on a growing body of literature on removing undesired information from text representations. This line of research has focused on developing algorithms to learn representations that are independent of protected attributes like demographic variables (Li et al., 2018; Raff and Sylvester, 2018; Barrett et al., 2019; Belrose et al., 2023). These methods typically leverage adversarial training or projection techniques to encourage models to learn representations orthogonal to the forbidden variable.

In addition, some work has also used LLMs for manipulating text, sometimes with goals related to disentanglement in mind. For example, work has used language models for style transfer and controlled text generation, aiming to modify attributes like sentiment or formality while preserving core content (Mir et al., 2019; Malmi et al., 2020). Other research has leveraged language models for tasks like paraphrasing (Krishna et al., 2020), simplification (Martin et al., 2020), and neutralizing biased language through LLM fine-tuning (Ghanbarzadeh et al., 2023). We build on these works by exploring the use of out-of-the-box language models for the task of forbidden variable removal via text disentanglement. To the best of our knowledge, we are the first to study the effectiveness of LLM prompting for the disentanglement task systematically and to compare this approach to human performance.

3 Defining Disentanglement

Previous work has focused on removing a forbidden variable from a numerical text representation, not from the raw text itself. For a representation X and a forbidden variable Z , Ravfogel et al. (2023) defined *concept erasure* as a process that finds a *guarding function* h such that $h(X) \perp Z$. Several approaches have been proposed to find guarding functions; most work has focused on finding projections that optimize guardedness with respect to linear classifiers (Ravfogel et al., 2020; Belrose et al., 2023). However, this process is not interpretable in that a human cannot easily reason about what the transformation is doing in the numerical representation space. In addition, finding the guarding function requires a collection of annotated training instances.

In this article, we use a similar conceptualization, but we transform the text directly instead of working in a latent representation space. We define *text*

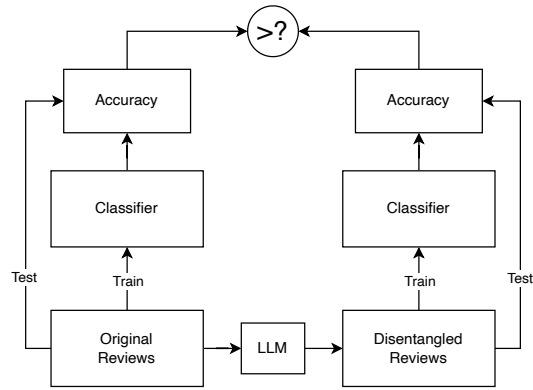


Figure 1: The experimental setup for measuring the effectiveness of LLMs at removing a target variable from the raw text representation.

disentanglement with respect to a forbidden variable Z as finding a guarding function d that takes a text and returns a transformed text where textual traces of Z have been removed so that $d(W) \perp Z$.

We can trivially satisfy this independence criterion by letting d return a transformed text that is unrelated to the original text W (e.g., an empty text). To avoid such uninformative cases, we also want the transformation to be minimally intrusive. In previous work that relied on linear projections, this condition was satisfied by construction.

Although the meaning of intrusiveness will vary with research context, we here define it as our ability to measure the effect of disentanglement on our ability to predict other variables represented in the text. Looking ahead, we envision that minimal intrusiveness can be defined on a semantic level in terms of maximizing similarity in a representation space or on a superficial level by minimizing string edit distance; at this point, we leave this formalization to future work.

4 Method

The goal of our experiments is to determine whether current out-of-the-box LLMs are able to disentangle a forbidden variable from text without removing traces of other variables. We test two LLMs: Mistral 7B and GPT-4. We choose Mistral 7B as a smaller open-source LLM that performs well on benchmarks and can be run on consumer hardware (Jiang et al., 2023). We choose GPT-4 as a top-performing commercially available model (as of 2024) (Achiam et al., 2023). More specifically, we use Mistral 7B v0.2 (instruction-tuned) and “gpt-4-0125-preview” from the OpenAI API.

To add context to our results, we also test the

performance of mean projection (Haghighatkhan et al., 2022), a method for removing specific information from representations, as well as, critically, the performance of human annotators. We include human annotators in our experiments to provide a benchmark for the level of forbidden variable removal that can be achieved by careful manual text editing, and to explore the intrinsic challenges that may stem from the entanglement of the forbidden variable and text content in natural language.

4.1 Dataset

In the experiments, we use a dataset consisting of 2000 Amazon reviews. This is a subset of the dataset published by Blitzer et al. (2007), which was originally created to investigate domain shifts and domain adaptation of sentiment classifiers. In our subset, each review has a label for the sentiment (positive or negative) and a label for the topic or product category (book, music, camera, health, DVD, or software). The dataset is approximately balanced in both sentiment and topic labels (e.g., 50.65% of reviews have a positive sentiment; the topic category with the largest number of reviews is “camera” with 17.8%; the category with the smallest number is “software” with 15.15%). We choose the sentiment as the forbidden variable and the topic as a proxy for the traces of other variables in the text we wish to keep.

We used this data because, in these reviews, sentiment information tends to be spread throughout the text, rather than localized to a specific sentence or section. This makes sentiment a challenging variable to remove while preserving other text content. As a result, this corpus serves as a strong test case—if LLMs (or humans) can successfully disentangle sentiment from these reviews while retaining the topic information—we would have compelling evidence of their text disentanglement capabilities. In other words, this test is strong in that we expect sentiment information to be spread throughout the text, as opposed to localized in a specific portion of the text, rendering the disentanglement task here more difficult than in the localized case.

4.2 Prompting

In order to instruct LLMs to disentangle a forbidden variable from text, we write the task description as a prompt. Given that the sentiment was chosen as the forbidden variable, the LLM is explicitly told to remove the sentiment from the reviews by revising them to be neutral. The LLM is also instructed

Rewrite the review such that the sentiment is completely neutral. It is very important that one cannot tell whether the review is positive or negative at all. Try and keep all other information in the review.

Here are a few examples of how to do this.

Example 1: [...]; Example 2: [...]; Example 3: [...]

Here's the review: \$Review

Figure 2: Excerpt from the few-shot prompt template. In our tests, \$Review is replaced with the original text of each review.

to keep all other information when revising the reviews. The LLM is not explicitly instructed to keep information about topic, as we want topic to be a proxy for all non-sentiment semantic information.

We investigate two prompting strategies for constructing prompts: few-shot prompting and prompt chaining. Few-shot prompting involves giving one or more examples of how to solve the task as part of the prompt. Doing this has been shown to improve model performance over a zero-shot setup and is one of the most common prompting techniques (Brown et al., 2020). Figure 2 shows an excerpt of the few-shot prompt used in our experiments. The full prompt is shown in Figure 7 in Appendix A.2. We provide 3 task examples to the LLM.

We also investigate prompt chaining as an alternative prompting strategy (Wu et al., 2022). Prompt chaining involves breaking the problem into smaller tasks and asking the LLM to complete the tasks one at a time, keeping the prompts and answers from previous tasks in context. This strategy can boost performance and gives access to intermediate reasoning steps, which can be used to better understand how the model reasoned about the task.

For our purposes, we use a 2-stage prompt chain. In stage 1, the LLM is asked to return a list of the parts of the text that seem to be associated with the forbidden variable. We include 3 examples of how to solve the task as part of the prompt for the first stage. In stage 2, the LLM is asked to rewrite the review from the first stage such that all traces of the forbidden variable are removed. Like before, the LLM is also instructed to keep parts of the review not associated with the forbidden variable. The full prompts for both stages are available in Figure 8 and Figure 9 Appendix A.2.

To investigate the default rewriting behavior of the LLMs we also test a paraphrase prompt, which

Setting	Prompt	Sentiment Accuracy ↓	Topic Accuracy ↑
No disentanglement		0.885 ± 0.035	0.946 ± 0.026
Mean projection		0.524 ± 0.054	0.946 ± 0.026
Human*	Prompt chaining	0.800 ± 0.145	0.842 ± 0.165
	Paraphrase	0.891 ± 0.037	0.951 ± 0.024
Mistral 7B	Few-shot	0.877 ± 0.023	0.951 ± 0.015
	Prompt chaining	0.841 ± 0.039	0.953 ± 0.023
	Paraphrase	0.899 ± 0.034	0.951 ± 0.024
GPT-4	Few-shot	0.824 ± 0.045	0.955 ± 0.024
	Prompt chaining	0.757 ± 0.044	0.945 ± 0.023

Table 1: The impact of disentangling text on sentiment and topic classification accuracy. Results are computed over 2000 Amazon reviews, except for the human setting which was computed on 152 reviews. ↑ and ↓ indicate whether higher values are or are not preferred.

asks the model to paraphrase the review without changing the meaning (see Figure 6 in Appendix A.2). If the LLM perfectly paraphrases the text without altering its semantic content, we would expect the rewritten text to exhibit the same level of association with the forbidden variable as the original text. However, if the LLM introduces any changes or artifacts during the rewriting process, this could amplify or diminish the signal of the forbidden variable compared to the initial text.

Finally, we also compare performance with a human baseline. In this baseline, we instructed three people via prompt chaining to (a) list out portions of the text related to sentiment and (b) re-write the text removing those sentiment-related portions but retaining all other information.

We make the raw reviews, the human-coded sentiment-related content, and the human-disentangled text available at doi.org/10.7910/DVN/TEC1ZP with replication repository at [GitHub.com/AIandGlobalDevelopmentLab/TextDisentanglement](https://github.com/AIandGlobalDevelopmentLab/TextDisentanglement).

4.3 Evaluation Design

We use two classifiers to evaluate the effectiveness of LLMs at disentanglement. Both classifiers are trained to label the sentiment of the reviews as this was our chosen forbidden variable. The first classifier is trained and tested on the original reviews; the second is trained and tested on the processed reviews. We then compare the accuracy of the two. Given that our dataset is approximately balanced, if the traces of the forbidden variable were successfully removed, we would expect the second classifier to have coin-toss accuracy. If traces of the forbidden variable were not removed, we would expect the two classifiers to have similar accuracy.

We use the same setup to test whether information about the topic is kept in the processed reviews. See Figure 1 for a summary of our evaluation setup.

We use logistic regression trained on document embeddings for all classifiers. The document embeddings are generated by taking the mean of all token embeddings within the document. Token embeddings are generated with DistilBERT (Sanh et al., 2019). To generate confidence intervals we bootstrap over the logistic regressions, using 500 bootstrap samples and a confidence level of 95%; we use an 80/20 train/test split.

5 Results

Our results are summarized in Table 1. In general, they show that current out-of-the-box LLMs are unable to consistently remove sentiment from data such as Amazon reviews. Of the LLMs and prompting strategies tested, GPT-4 with prompt chaining achieves best (i.e., lowest) sentiment classifier accuracy with an average of 75.7% (see an example in Appendix A.1). While the human annotators also struggled with removing traces of the sentiment from the reviews, they achieve results comparable with GPT-4 in sentiment accuracy. The mean projection experiment shows that removing almost all traces of sentiment from the reviews while keeping traces of the other variables is possible when operating at the representation level.

The LLMs performed well at keeping information about topic. Furthermore, the paraphrase prompting strategy shows that they successfully keep information about sentiment when rewriting: in the case of GPT-4, the prompt even leads to a slight increase in sentiment classifier accuracy. This indicates that LLMs may amplify original text

signals when paraphrasing.

6 Implications

While the LLMs generally struggle to remove statistical associations with the forbidden variable, we also find that human coders face a similar difficulty. This finding implies that there may be, at least for the task described here, limited separability in the text between the forbidden variable and the remainder of the text. This limited separability raises questions about the robustness of disentanglement methods operating on downstream representations as such methods may inadvertently generate representations incongruous with real text. Further research is therefore needed to develop techniques that can effectively disentangle the forbidden variable from the relevant semantic content in a way that respects the content of the original text.

Ethical Considerations

Our work explores the capabilities and limitations of large language models and human annotators in disentangling text variables, raising several important ethical considerations. For example, our findings highlight the challenges of completely removing traces of a target variable while preserving other semantic content in the raw text space. This underscores the importance of transparency and interpretability when applying this or other disentanglement methods, as residual signals may still be detectable even after processing or the processed text representations may be incongruent with the original text semantics. It is important to communicate these limitations to end-users and decision-makers.

Limitations

Our experiments focus on variables where the relevant information is spread throughout the text, such as sentiment in product reviews. However, in some applications, the forbidden variable may be more localized and separable, such as personal information like names or addresses (Hovy and Prabh-moye, 2021). In these cases, the disentanglement task may be easier, as the target information can be more precisely removed.

We also focus on sentiment and topic variables that are relatively independent in our dataset. However, in real-world scenarios, variables of interest may be more intrinsically interrelated, such as political ideology and slant in news articles. Disentan-

gling inherently correlated variables while preserving salient information could be more challenging (Daoud et al., 2022).

Finally, our evaluation results rely on machine learning classifiers, which may not fully capture human perception of the removal of the forbidden variable. Classifiers detect statistical patterns but do not necessarily “read” text like humans do. There could be cases where classifiers detect residual signals that are not semantically meaningful to humans or where important nuances are lost that are more apparent to humans. To get a more complete picture of disentanglement effectiveness, future work should augment machine evaluations with human judgment experiments, such as having annotators guess the original target variable from the disentangled text.

Acknowledgements

This research was supported by the project *Countering Bias in AI Methods in the Social Sciences* under the Wallenberg AI, Autonomous Systems and Software Program – Humanity and Society (WASP-HS), funded by the Marianne and Marcus Wallenberg Foundation and the Marcus and Amalia Wallenberg Foundation.

The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at Alvis partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. [GPT-4 technical report](#). *arXiv preprint arXiv:2303.08774*.
- Maria Barrett, Yova Kementchedjheva, Yanai Elazar, Desmond Elliott, and Anders Søgaard. 2019. [Adversarial removal of demographic attributes revisited](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6330–6335, Hong Kong.
- Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. [LEACE: Perfect linear concept erasure in closed form](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 66044–66063.

- John Blitzer, Mark Dredze, and Fernando Pereira. 2007. [Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification](#). In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic.
- Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. [Man is to computer programmer as woman is to home-maker? Debiasing word embeddings](#). In *Advances in Neural Information Processing Systems*, volume 29.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. [Language models are few-shot learners](#). *Advances in neural information processing systems*, 33:1877–1901.
- Adel Daoud, Connor Jerzak, and Richard Johansson. 2022. [Conceptualizing treatment leakage in text-based causal inference](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5638–5645, Seattle, United States.
- Yaroslav Ganin and Victor Lempitsky. 2015. [Unsupervised domain adaptation by backpropagation](#). In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France. PMLR.
- Somayeh Ghanbarzadeh, Yan Huang, Hamid Palangi, Radames Cruz Moreno, and Hamed Khanpour. 2023. [Gender-tuning: Empowering fine-tuning for debiasing pre-trained language models](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5448–5458, Toronto, Canada.
- Lin Gui and Victor Veitch. 2023. [Causal estimation for text data with \(apparent\) overlap violations](#). In *Proceedings of the Eleventh International Conference on Learning Representations*, Kigali, Rwanda.
- Pantea Haghighatkhah, Antske Fokkens, Pia Sommerauer, Bettina Speckmann, and Kevin Verbeek. 2022. [Better hit the nail on the head than beat around the bush: Removing protected attributes with a single projection](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8395–8416, Abu Dhabi, United Arab Emirates.
- Yuzi He, Keith Burghardt, and Kristina Lerman. 2020. [A geometric solution to fair representations](#). In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 279–285, New York, NY, USA. Association for Computing Machinery.
- Dirk Hovy and Shrimai Prabhumoye. 2021. [Five sources of bias in natural language processing](#). *Language and Linguistics Compass*, 15(8):e12432.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. [Mistral 7B](#). *arXiv preprint arXiv:2310.06825*.
- Katherine Keith, David Jensen, and Brendan O’Connor. 2020. [Text and causal inference: A review of using text to remove confounding from causal estimates](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5332–5344, Online.
- Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. [Reformulating unsupervised style transfer as paraphrase generation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online.
- Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. [Towards robust and privacy-preserving text representations](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 25–30, Melbourne, Australia.
- Eric Malmi, Aliaksei Severyn, and Sascha Rothe. 2020. [Unsupervised text style transfer with padded masked language models](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8671–8680, Online.
- Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. [Controllable sentence simplification](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France.
- Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. [Evaluating style transfer for text](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 495–504, Minneapolis, United States.
- Edward Raff and Jared Sylvester. 2018. [Gradient reversal against discrimination: A fair neural network learning approach](#). In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 189–198.
- Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. [Null it out: Guarding protected attributes by iterative nullspace projection](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online.
- Shauli Ravfogel, Yoav Goldberg, and Ryan Cotterell. 2023. [Log-linear guardedness and its implications](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9413–9431, Toronto, Canada.

- Margaret E. Roberts, Brandon M. Stewart, and Richard A. Nielsen. 2020. [Adjusting for confounding with text matching](#). *American Journal of Political Science*, 64(4):887–903.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. [DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter](#). *arXiv preprint arXiv:1910.01108*.
- Petter Törnberg. 2024. [Best practices for text annotation with large language models](#). *arXiv preprint arXiv:2402.05129*.
- Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. [AI chains: Transparent and controllable human-AI interaction by chaining large language model prompts](#). In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, pages 1–22, New York, NY, USA. Association for Computing Machinery.
- Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. [Can large language models transform computational social science?](#) *Computational Linguistics*, pages 1–55.

A Appendix

A.1 disentanglement Example

i bought this item because my husband heard is the best internet security . now i regret buying one . since ive installed it , it slowed my computer , its very very hard to close window and worse when i disconnected from the internet i have a hard time logging off . everytime im on the internet , i get mad and upset . its not worth all the stress . i definitely dont recommend this product . now im trying to figure how i can return the product and go back to norton

Figure 3: Example of an original review from the Amazon dataset

The parts of the review that contain information about the sentiment are:

- * my husband heard is the best internet security
- * now i regret buying one
- * it slowed my computer
- * its very very hard to close window
- * worse when i disconnected from the internet i have a hard time logging off
- * everytime im on the internet, i get mad and upset
- * its not worth all the stress
- * i definitely dont recommend this product
- * now im trying to figure how i can return the product

Figure 4: GPT-4 response for the first stage of prompt chaining with the review from Figure 3

I purchased this item after hearing it was considered a top choice for internet security. Since installing it, I have observed changes in my computer's performance and behavior when disconnecting from the internet, as well as the process required to log off. I am currently exploring options to return the product and considering switching back to Norton.

Figure 5: GPT-4 response for the second stage of prompt chaining with the review from Figure 3 and the first stage response from Figure 4

A.2 Prompts

Paraphrase the following text without changing the meaning:

\$REVIEW

Figure 6: Prompt template for the “identity” experiments.

Rewrite the review such that the sentiment is completely neutral. It is very important that one cannot tell whether the review is positive or negative at all. Try and keep all other information in the review.

Here are a few examples of how to do this.

Example 1: if the original review was:

i bought this album because i loved the title song . it 's such a great song , how bad can the rest of the album be , right ? well , the rest of the songs are just filler and are n't worth the money i paid for this . it 's either shameless bubblegum or oversentimentalized depressing tripe . kenny chesney is a popular artist and as a result he is in the cookie cutter category of the nashville music scene . he 's gotta pump out the albums so the record company can keep lining their pockets while the suckers out there keep buying this garbage to perpetuate more garbage coming out of that town . i 'll get down off my soapbox now . but country music really needs to get back to it 's roots and stop this pop nonsense . what country music really is and what it is considered to be by mainstream are two different things .

then the neutral rewrite might be:

I bought this album because of the title song. The rest of the album I didn't know as well. Kenny Chesney is a popular artist in the Nashville music scene. He makes many albums with his record company. Country music has been evolving from its roots to a more pop sound.

Example 2: if the original review was:

this is a very good shaver for the private area . however , the key to getting the best results is to trim the longer hairs with scissors or the largest guard first . this will keep the shaver from pulling on the longer hairs and will enable the foil part of the shaver to work . the foil will not be able to do its job if the hairs are too long . the only problem i had with the shaver was that it did not enable me to shave my back like it claimed . however , i use the ' mangroomer ' back shaver for this and it is perfect for you to shave off all your back hair easily with its elongated handle . it is a great product as well . therefore , i would have to say these two products coupled together seem to cover all the bases for men 's grooming on the body . i would highly recommend both of them for perfect manscaping results

then the neutral rewrite might be:

To use this shaver in the private area it is important to trim the longer hairs with scissors or the largest guard first. This will keep the shaver from pulling on the longer hairs and will enable the foil part of the shaver to work. The foil will not be able to do its job if the hairs are too long. The shaver might also not work well on the back. For this, there are other options such as the 'Mangroomer' back shaver which has an elongated handle that makes it easy to shave back hair.

Example 3: if the original review was:

i bought bead fantasies and bead fantasies ii at the same time after reading the positive reviews ; i wish i had looked at these books before buying . there are pretty motifs that i will incorporate into my beading projects but i find the small typed directions overly simplistic and the diagrams are too small . i 'm glad this is n't my first beading book or i would feel totally discouraged from trying any of these projects . i wo n't be buying bead fantasies iii . the art and elegance of beadweaving and coraling technique remain my favorite beading books .

then the neutral rewrite might be:

I bought Bead Fantasies and Bead Fantasies II at the same time. I like some of the motifs but not others. This is not my first beading book. The art and elegance of beadweaving and coraling technique are great beading books.

Here's the review:

\$REVIEW

Figure 7: Prompt template for the “few-shot“ experiments.

Identify the places in the following review which contain information about the sentiment and return them as bullet points.

Here are a few examples of how to do this.

Example 1: if the original review was:

i bought this album because i loved the title song . it 's such a great song , how bad can the rest of the album be , right ? well , the rest of the songs are just filler and are n't worth the money i paid for this . it 's either shameless bubblegum or oversentimentalized depressing tripe . kenny chesney is a popular artist and as a result he is in the cookie cutter category of the nashville music scene . he 's gotta pump out the albums so the record company can keep lining their pockets while the suckers out there keep buying this garbage to perpetuate more garbage coming out of that town . i 'll get down off my soapbox now . but country music really needs to get back to it 's roots and stop this pop nonsense . what country music really is and what it is considered to be by mainstream are two different things .

then the parts of the review that contain information about the sentiment are:

- * i loved the title song
- * it 's such a great song
- * the rest of the songs are just filler and are n't worth the money
- * it 's either shameless bubblegum or oversentimentalized depressing tripe
- * the suckers out there keep buying this garbage to perpetuate more garbage coming out of that town
- * but country music really needs to get back to it 's roots
- * nonsense

Example 2: if the original review was:

this is a very good shaver for the private area . however , the key to getting the best results is to trim the longer hairs with scissors or the largest guard first . this will keep the shaver from pulling on the longer hairs and will enable the foil part of the shaver to work . the foil will not be able to do its job if the hairs are too long . the only problem i had with the shaver was that it did not enable me to shave my back like it claimed . however , i use the ' mangroomer ' back shaver for this and it is perfect for you to shave off all your back hair easily with its elongated handle . it is a great product as well . therefore , i would have to say these two products coupled together seem to cover all the bases for men 's grooming on the body . i would highly recommend both of them for perfect manscaping results

then the parts of the review that contain information about the sentiment are:

- * this is a very good shaver for the private area
- * the only problem i had with the shaver was that it did not enable me to shave my back like it claimed
- * it is perfect for you to shave off all your back hair easily with its elongated handle
- * it is a great product as well
- * i would highly recommend

Example 3: if the original review was:

i bought bead fantasies and bead fantasies ii at the same time after reading the positive reviews ; i wish i had looked at these books before buying . there are pretty motifs that i will incorporate into my beading projects but i find the small typed directions overly simplistic and the diagrams are too small . i 'm glad this is n't my first beading book or i would feel totally discouraged from trying any of these projects . i wo n't be buying bead fantasies iii . the art and elegance of beadweaving and coraling technique remain my favorite beading books .

then the parts of the review that contain information about the sentiment are:

- * i wish i had looked at these books before buying
- * there are pretty motifs
- * i find the small typed directions overly simplistic
- * the diagrams are too small
- * i 'm glad this is n't my first beading book
- * i would feel totally discouraged
- * i wo n't be buying bead fantasies iii

Here is the review:

\$REVIEW

Figure 8: Prompt template for the first stage of the “prompt chaining” experiments.

Rewrite the original review such that all the information identified about the sentiment is removed. The goal is to make the review completely neutral. It is very important that one cannot tell whether the review is positive or negative at all. Keep all other information in the review.

Figure 9: Prompt template for the second stage of the “prompt chaining” experiments.

Retrieval Augmented Generation of Subjective Explanations for Socioeconomic Scenarios

Razvan-Gabriel Dumitru^{†*} Maria Alexeeva^{†*} Keith Alcock[†] Nargiza Ludgate[‡]
Cheonkam Jeong[†] Zara Fatima Abdurahaman[◁] Prateek Puri[◁]
Brian Kirchoff[◊] Santadarshan Sadhu[◊] Mihai Surdeanu[†]

[†] University of Arizona, Tucson, AZ, USA [‡] University of Florida, Gainesville, FL, USA

[◁] RAND Corporation, Santa Monica, CA, USA

[◊] NORC at the University of Chicago, Chicago, IL, USA

{rdumitru, alexeeva, msurdeanu}@arizona.edu

Abstract

We introduce a novel retrieval augmented generation approach that explicitly models causality and subjectivity. We use it to generate explanations for socioeconomic scenarios that capture beliefs of local populations. Through intrinsic and extrinsic evaluation, we show that our explanations, contextualized using causal and subjective information retrieved from local news sources, are rated higher than those produced by other large language models both in terms of mimicking the real population and the explanations quality. We also provide a discussion of the role subjectivity plays in evaluation of this natural language generation task.

1 Introduction

Retrieval augmented generation (RAG) has emerged as a powerful technique to mitigate the limited and static knowledge horizon of large language models (LLMs) (Lewis et al., 2020; Guu et al., 2020). However, RAG methods struggle with tasks that cannot easily be captured through search (Yan et al., 2024; Asai et al., 2024). For example, DARPA’s Habitus program¹, which aims to ingest subjective information from local populations into scientific models, recently organized an evaluation, descriptively called *Predict what the Locals would Predict* (PWLWP), in which natural language generation (NLG) systems had to predict the responses of a local population to several hypothetical socioeconomic scenarios. In particular, the population of interest consisted of adults from the Ashanti region of Ghana; all scenarios focused on mining in the region. An example of the such a scenario and an explanation by our approach is shown in Table 1.

To properly address this task, this work is the first to propose a RAG approach that explicitly

models causality (so we can generate causal explanations) and subjectivity (so we can capture the beliefs of a local population).² The key contributions of our work are:

(1) A RAG method that incorporates causality and subjectivity. In particular, our RAG method contains two additional components that detect causality and subjectivity in the retrieved sentences; we add only such meaningful sentences to the context of the downstream LLM.

(2) A first PWLWP evaluation in which we show that: (a) a RAG-extended LLM provides more accurate responses than a “vanilla” one that operates without context; and (b) the LLM that has access to sentences that contain subjective descriptions of causal phenomena scores higher (both in terms of mimicking the real population, and in the quality of explanations) than other LLMs.

(3) An analysis and discussion of explanation quality. Unsurprisingly, the evaluation of subjectivity is subjective itself. We discuss the different perspectives emphasized by different annotators and why annotation differences should be encouraged in this task (Plank, 2022).

2 Approach

In this section, we provide a brief overview of the system, with the pipeline described in detail in Section 3 and the experiment in Section 4.

2.1 The task

In this paper, we generate detailed explanations for a set of socioeconomic scenarios related to mining in Ghana. The task was administered by the RAND Corporation think tank as a benchmark evaluation for the Predicting What Locals Would Predict (PWLWP) portion of the *DARPA Habitus*

*equal contribution

¹<https://www.darpa.mil/program/habitus>

²Resources for this work are available at <https://github.com/clulab/habitus/wiki/NLP-CSS-Workshop-2024>

scenario	Imagine that China has recently completed a mining infrastructure project in Brong Ahafo, Ghana. Chinese gamblers in the region have received an offer to work in the new mine. Additionally, imagine that the Chinese government has promoted the new large-scale mining effort as a source of employment for locals engaged in small-scale illegal mining. Suppose that within a month, the Ghanaian government has seen a decrease in illegal mining activities in Brong Ahafo. This is because...
generation	... The promise of formal employment in a large-scale mining operation can be seen as a powerful incentive for Ghanaians to cease their participation in illegal mining, especially since "74% of Ghanaians say no citizen should be permitted to engage in illegal small scale mining or "gambsey" for any reason". Offering an alternative, legal prospect for income can prod these individuals to abandon illegal mining...

Table 1: Excerpt from an explanation generated for one of the test scenarios using sentences that contain locals’ beliefs as context passed to the model in the prompt.

program. As its name indicates, this task aims to emulate the thinking of local populations without interviewing people on the ground. Each scenario consists of a description of a situation and its result, followed by a multiple choice question. A sample scenario is in Table 2 of the Appendix.

2.2 Generating predictions and explanations

Our RAG method outputs the probabilities of the multiple choice responses for each scenario and produces detailed explanations for each question by using thematically-related retrieved sentences. We prompt the model to make use of and cite contextual data to provide support for the explanations. For contextual data, we use sentences extracted from online news articles pertaining to the subject of interest, mining, and served by Ghanaian media outlets. Context data containing several types of information, notably beliefs of the local communities and causal relations, is intended to provide the model with the location-specific knowledge that it may not have access to. Additionally, by providing causal and subjective information, we hope to improve the model’s explanatory power.

2.3 Evaluation

We perform two types of evaluation: (a) we calculate the accuracy of our approach in answering the multiple choice questions accompanying each scenario (a sample question and answer choices are shown in Table 2), and (b) evaluate the quality of the explanations that it generates in response to each scenario. The gold data for the accuracy calculation comes from a survey conducted among the local population (1,782 households) at the target location (the Ashanti region of Ghana) by the research organization NORC at the University of Chicago. The survey methodology is discussed in Appendix B.

For the explanations evaluation, we ask two types of evaluators—domain experts and linguists—

to evaluate the generated explanations by providing a score and the rationale for the score. Based on these, we compare the explanations generated using different context types and devise a set of evaluation criteria to use for this type of NLG task.

3 System Overview

In this section, we provide the technical details of the approach used for the task. As mentioned above, our approach is an instance of retrieval augmented generation, expanded to rely on context that is relevant for the task, i.e., context sentences that are likely to describe beliefs held by a local population for a given hypothetical scenario. The overview of the proposed architecture is shown in Figure 1. We discuss the key contributions below.

3.1 Data Sources

Retrieval of a sentence corpus began with the observation that remarkably many Ghanaian news articles are published online. A BBC media guide³ seeded a search for sites with links to the most prominent media outlets. Further investigation found sites absent of paywalls, with favorable terms of service, and, conveniently for our tooling, written in English, an official language of the country. We were able to identify at least one suitable representative each in categories for radio, television, the press, and news agencies.

A list of simple search terms related to the issue of mining was created and used initially to gauge the suitability of a site’s article collection, with bigger being better: mining, gold, gambsey (small-scale, illegal gold mining), harvest, livestock, crop, and price. Each site’s native search mechanism was employed, rather than the alternative site-specific Google search, to ensure that only news reports were returned. Although some false hits resulted

³<https://www.bbc.com/news/world-africa-13433793>

Scenario	Answer Choices
Imagine the Ghanaian government implements reforms that change the time it takes for local residents to obtain a legal mining license, reducing the time from three years to three months. Suppose that within three months, the number of mining license applications received by the government tripled. This would have been most likely because...	<ol style="list-style-type: none"> 1. Those involved in illegal mining have begun working for the Chinese mining company 2. Those involved in illegal mining have sought other opportunities outside of the region because of the Chinese mine opening 3. Those involved in illegal mining have sought opportunities in small-scale agriculture because of the Chinese mine opening 4. Those involved in illegal mining have sought other opportunities within the region unrelated to agriculture 5. None of the above

Table 2: Sample evaluation scenario with five answer choices.

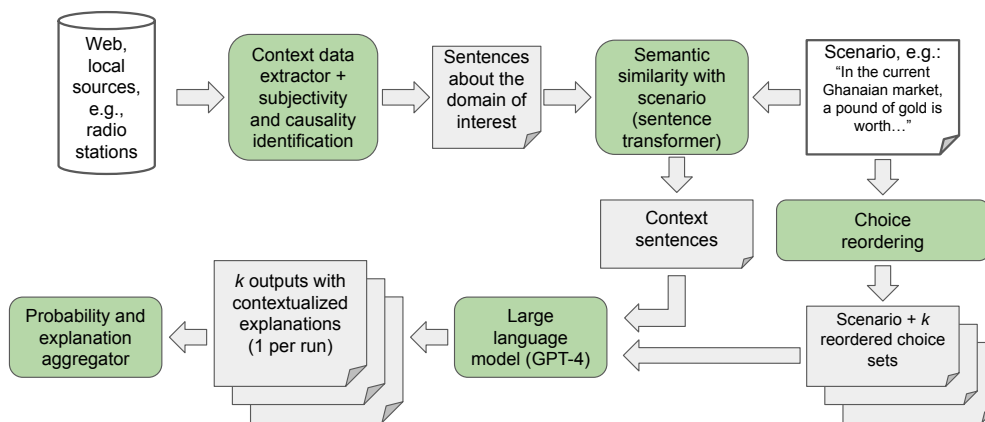


Figure 1: Overview of the natural language generation (NLG) pipeline: software components in green, input data in white, and intermediate data in grey.

from the simplicity of the searches, like “gold” unearthing sports articles, they were not filtered out at this stage because they are later accounted for when context is constructed.

For acceptably prolific sites, the multiple pages of hits generated by the queries were further processed into article lists, and listed articles were then downloaded and finally parsed by the `scala-scraper` library⁴ to identify article title, publication date, and byline for tracking provenance, and to assemble the article’s text with as much extraneous markup as possible removed. Approximately 70,000 news articles with publication dates ranging from 2013 to 2023 were collected.⁵ This resulted in a corpus of over 1.3 million sentences originating from seven sites (for the list of

the sites used, see Appendix C).

3.2 Classifying Subjective Information

As one type of background information, we extract sentences containing subjective views of local populations. These can be either beliefs, that is subjective views on how the world works, or attitudes, that is how people feel about something. For instance, the following sentence contains a belief (in bold) held by a subset of the population in Ghana:

*The project manager also pointed out that three-fourths (74%) of Ghanaians say **no citizen should be permitted to engage in illegal small-scale mining or "galamsey" for any reason** ...*

To identify sentences that contain subjective views reportedly held by local populations (subsequently just “beliefs” for brevity), we use the binary classifier described in (Alexeeva et al., 2023). We run the fine-tuned BERT-based model on text of

⁴<https://github.com/ruippeixotog/scala-scraper>

⁵The list of URLs is available at <https://github.com/clulab/habitus/wiki/NLP-CSS-Workshop-2024>

all the documents retrieved in the previous step and use the sentences that were classified as containing beliefs with a confidence over 0.97.

3.3 Identifying Causality

We identify sentences containing causal relations by using Eidos, the machine reading library focusing on extraction of causal statements from text (Sharp et al., 2019). A causal relation is a binary cause-effect relation between two concepts, with one influencing the other in a positive (e.g., promotion) or negative (e.g., inhibition) way. Based on our analysis of 50 sentences identified as containing causal relations, 76% indeed contained a relation of the intended type. Some sentences were judged to be false positives due to lack of specificity of the concepts involved (ex. a), or the causal relations identified indicate hypothetical scenarios or recommendations rather than factual information (ex. b):

- a. *This has resulted in a renewed public discussion on illegal mining activities.*
- b. *Ghana needs to ... encourage middle scale and large scale farming that would contribute immensely to total yield of agriculture produce to provide more food, employment and help reduce importation.*

We note that we do not verify whether the extracted relations are indeed causal, but accept them as such since they were extracted based on textual cues consistent with indicating causality.

Similarly to how we extract sentences with subjective views, we run the rule-based extraction system on the full text of documents retrieved for the given scenario and obtain a set of sentences that contain causal relations, e.g.:

Government intends to make conscious efforts to integrate the mining industry with the rest of the local economy, thus making it possible for Ghanaian entrepreneurs to increase their participation in the mining industry

3.4 Retrieving Thematically-Related Information

Considering that our sentence corpus is of significant size, we need a way to reduce it to a subset of sentences that: (a) is small enough to fit into the context size available to the LLM (e.g., 16K for GPT-4 at the time of our experiments), and (b) is relevant to the scenario at hand.

In order to make the most of the small context window available we used a sentence transformer, more specifically allMiniLM_L6_v2 (Tanner, 2023), which extracts sentences from our corpus that are semantically similar to the description of the scenario. To filter the sentences we first query the similarity score between each sentence in the corpus and the scenario to be tested (excluding the possible answer choices) and then we sort them according to that score. We decided to remove the choices from the similarity evaluator because they might introduce biases in the sentences retrieved.

When retrieving information we also make sure that all of the context sentences that we extract match the expected context type. Context type has five possible variations: all sentences, belief sentences, causal sentences, causal belief sentences (i.e., sentences containing subjective statements that include causality), and no context. The last setting is used to test a “vanilla” GPT-4, i.e., an LLM using a prompt without any contextual information.

Finally, we sort the sentences of the given type in descending order of their similarity to our input scenario. To build our context sentences, we pick sentences starting from the most similar until we sum of tokens for our context plus the rest of the prompt reaches the context token limit, making sure to fit in as much information as possible. Further, we prompt the LLM to use these sentences in its explanation and to cite them accordingly.

3.5 Retrieval Augmented Generation

In order to force GPT-4 to process the prompt before ranking the choices, we prompt the model in two independent steps. First we ask it to provide justifications for its decisions by citing information from the provided context sentences, and then we ask it to rank the choices. We observe that this approach improves the results, as GPT-4 will use the information it extracted to rank the choices, instead of directly ranking the choices. Next we detail each step and corresponding prompt.

First prompt:

Read the following question delimited with backticks:

```
1 ‘‘‘{scenario}’’‘
```

Use the following context sentences delimited with backticks as background knowledge:

```
1 ‘‘‘{context}’’‘
```

Provide long and thorough justifications for each of the choices independently, without referring to the other choices, while citing the context using quotes:

```
1 ‘ ‘{choices}‘ ‘
```

where $\{scenario\}$ refers to the question for which we want to rank choices. $\{context\}$ refers to the sentences that we use to answer the question. $\{choices\}$ refers to a list of choices from which the model can choose, which we further detail in the next subsection. The main goal of this prompt is to force GPT-4 to use the context and cite it correctly. We also make it justify each choice independently so that it has more information when it ranks all of the choices. For consistency, we kept the prompt unchanged even in the “contextless” scenario, which resulted in the model using the text of the scenario as citable context information.

Second prompt:

Rank each choice from most likely to be true to least likely and copy the justification as JSON format with the fields:

```
1 (id, choice, rank, justification)
```

This prompt ensures that the information that we generated before is now used to correctly rank the choices, as well as enforcing a JSON format.

3.6 Choice Reordering and Probability and Explanation Aggregation

In initial experiments, we observed a slight bias in that GPT-4 is more likely to rank the first choices given in its prompt as being better. To alleviate this we decided to roll the choices in all possible variations. We also had the option to permute the choices in all possible variations, while this would be more precise and it would break any bias between the choices’ ordering, it would lead to a very high computational demand for each query. For N possible choices, rolling the choices leads to N possible variations that need to be run, while permuting it in all possible ways would lead to $N!$ runs. In general, the set of ordering that we run is the following:

- Option 1: (1 2 3 ... n)
- Option 2: (2 3 4 ... 1)
- Option 3: (3 4 5 ... 2)
- ⋮
- Option n : (n 1 2 ... n - 1)

The last problem that we need to solve is how we aggregate all of those runs into a final set of probabilities. To this end, we first invert each rank so that higher values are positive. For example, if GPT-4 ranks a choice as number 1 and we have 5 choices, we convert the rank to $choices - rank$ meaning $5 - 1 = 4$ in our case. Next we sum up all of the ranks obtained for each of the N runs with different orderings into a single vector denoted $final_rankings$ that has one value per choice. To transform this vector into probabilities, we used softmax with a tuned gamma parameter (to avoid overly peaked distributions):

$$P(\text{choice}_i) = \frac{e^{\gamma \cdot \text{final_rankings}(\text{choice}_i)}}{\sum_{\text{rank} \in \text{final_rankings}} e^{\gamma \cdot \text{rank}}}$$

For a walk-through example, see Appendix A.

4 Experiment

For the experiment, we produce answers to the multiple choice questions and generate explanations for seven scenarios. For each scenario, we produce generations using five different types of context provided to the model: no context and context in the form of thematically related sentences containing either beliefs, causal relations, causal beliefs, or just information related to the topic. See Table 4 in the Appendix for excerpts of generations produced with each type of context.

For the evaluation of our response rankings in relation to the local population survey gold data (see Appendix B for survey methodology), we compare the probability distribution for all choices produced by our system with the distributions produced by three baselines using mean absolute error (MAE). The baselines are LLM-based with some additional features for information retrieval, chain of thought prompting, and inclusion of population identity context. Two of the baselines use GPT-4 and differ in how they produce distributions to compare against the survey data: the ‘GPT-4 TopVote’ baseline does that by keeping track of how many times each multiple choice answer was selected as the top choice over a number of samples; and the ‘GPT-4 Calibrated’ baseline assigns weights to each response based on the answer choice ranking produced for each sample. The third baseline uses an offline LLM (Mistral 7B), simplified to approximate a real world scenario, where resources may be limited and privacy concerns may preclude the use of online LLM APIs. For benchmark implementation and metrics details, see Appendix D.

For the explanations evaluation, we assemble the scenarios and the generated explanations into a spreadsheet, with one scenario per row followed by the explanations presented in random order. The generated explanations come in a standard essay format, with an introduction, between two and six body paragraphs, and a conclusion. We ask the annotators to provide a score for each explanation without knowing the context type setting on a scale from 0 to 10 (ten being the highest) and provide a rationale for their score.

The evaluation was done by two annotators: a domain expert and a linguist. Additionally, an annotation supervisor (another linguist) supervised the annotation process and provided meta analysis of the rationale and additional comments on the quality of the outputs.

While some authors point out various issues with NLG evaluation criterion inconsistency (Gehrmann et al., 2023; Howcroft et al., 2020), we chose not to provide any specific evaluation guidelines to the evaluators. With the evaluation being closely connected to a project in a real world setting, it was crucial to see what criteria the domain expert views as relevant for their field in this evaluation, and we did not expect previously defined, non-domain-specific criteria to be necessarily relevant. For the linguist evaluator, we expected the criteria to be similar to those described in previous work, but we chose to not provide detailed guidelines beyond a few example criteria to keep the evaluation procedure consistent between the two annotators.

5 Results and Discussion

5.1 Evaluation of Probability Distributions for the Multiple Choices

Figure 2 shows the MAE boxplot distribution of two versions of our system and three baselines on the task of approximating the distribution of the answers provided by the local population based on the NORC survey. For this evaluation, we used an LLM with no context and the context of thematically-related sentences. Our best system (“Thematically-related”) performs at the same level as the *GPT-4 Calibrated* baseline, while, more importantly for our aim, also providing high quality explanations (see subsequent sections). We find it encouraging that adding context for explanations improves the performance of the overall method, i.e., with the MAE for the context setting lower than the contextless one.

5.2 Impact of Background Information Types on Explanations

Figure 3 shows the mean scores for each scenario and context type based on the evaluation by two annotators (a domain expert and a linguist). As seen in the figure, the explanations generated with the use of context, that is instances of retrieval augmented generation, are scored higher by the annotators than those without the context provided. From this, we conclude that retrieval-augmented generation has the potential to help with generating explanations for local events. This could be because the background information that we feed to the model through the prompt provides the knowledge that is not well represented in the model.

Moreover, including subjective views as context (“beliefs” and “causal beliefs” settings in Figure 3) results in higher rated explanations. This could indicate that beliefs, along with causal relations, have a potential to explain functioning of complex systems, as discussed in (Alexeeva et al., 2023).

5.3 Evaluation Criteria

We observed that for several datapoints, the scores between the two annotators were quite different (see Figure 6 in the Appendix). By analyzing the rationale provided by the annotators for their scores, we identified two broad categories of evaluation criteria—content quality and text quality,—which could be prioritized differently by the two types of annotators (domain experts vs. linguists), thus resulting in differences in scoring.

5.3.1 Content Quality

Some of the more prominent content-related criteria mentioned by the annotators included the logic of the explanation, the number of factors contributing to the explanation, and task comprehension—that is whether or not the explanation was relevant to the prompt scenario.

For our experiment, the most interesting evaluation criterion, and also the one mentioned most by the annotators, was the use of quotations as evidence. Since we kept the prompt consistent in how the model was instructed to cite context, the settings both with and without background information resulted in generations that included extensive quotations; the only difference was what was being cited: the provided background information or the text of the scenario. This consistency in format of the output between the two setting categories makes this a reasonable comparison.

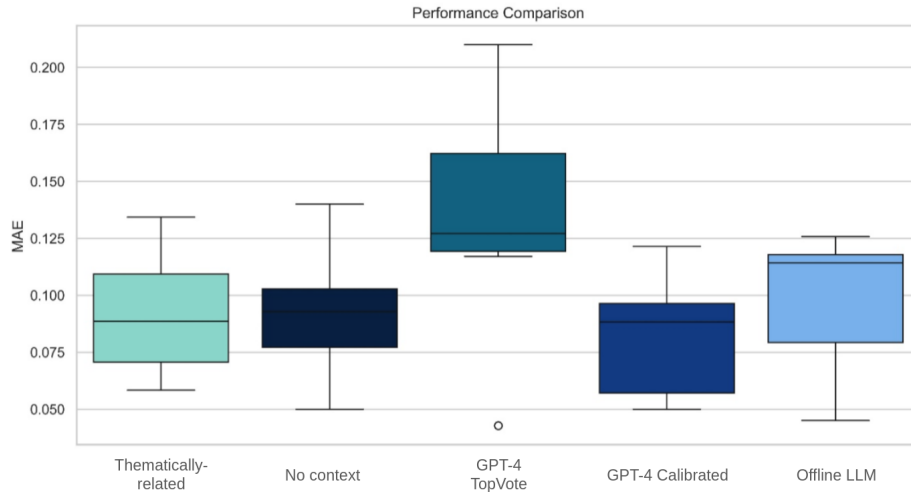


Figure 2: Mean absolute error (MAE) boxplot distribution (lower score is better). The performance of the two of our systems that were evaluated (“Thematically-related” and “No context”) is comparable to that of the strongest baseline (“GPT-4 Calibrated”) and better than two other baselines (“GPT-4 TopVote” and “Offline LLM”).

In terms of quality, the annotators were looking for the quotes to be relevant to the scenario and the topic of the paragraph, the number of quotes used, the quality of connection of the quote to surrounding text, the amount of elaboration on the quote (Ex. 1 in Table 6), as well as the quote being grounded to a source (e.g., attribution to a person).

An additional aspect of quotations use was their accuracy, that is whether or not the quotation came from the provided context sentences and whether it was modified by the model. This criterion was not brought up by the annotators since they did not have access to the full input that was provided to the model. The annotation supervisor performed a spot check of the quotes used in the generated explanations and did not find any instances where the quotations were inaccurate with the exception of one instance of parentheses being left out. Interestingly, there is some inconsistency in the length of quotes the model uses, with the length ranging from single words to full, multi-clause sentences. While single-word quotes are impressionistically more successful, implying synthesis of ideas instead of copying, they are harder to verify as they require careful rereading of the associated context sentences (Ex. 2 in Table 6).

Overall, the use of quotation was viewed as more efficient in the context-based settings than in the contextless one, although individual data points in both categories were viewed as using citations more or less efficiently, from being judged as insufficient to excessive.

5.3.2 Text Quality

Prominent factors related to text quality were style (presence of repetition, wordiness, and tone); organization (maintaining a standard essay-style structure of introduction, several body paragraphs, and a conclusion); and presentation of the output, for instance, whether or not the paragraphs describing various factors were numbered.

An interesting criterion for this task was the use of hedging, which one of the annotators used as a proxy for model confidence. The hypothesis was that a model would use less hedging when there is more evidence that it can provide, or, in other words, the less unsubstantiated reasoning (or “hallucinations”) the the model needs to output, the fewer hedges it will use. While for most settings, the annotations on the use of hedging were inconsistent (e.g., three causal outputs were judged as having a high level of hedging and three as low), no-context setting outputs were mainly judged as being hedging-heavy and belief context outputs as minimal or moderate in use of hedging.

Another key criterion is the match between various components of the generated text. Based on this criterion, issues can come up on different levels: the generated explanation might not match the scenario in the prompt in terms of content or style; a paragraph, while sensible on its own, may not match the thesis statement of the introduction or may not be logically connected to the preceding paragraph; a quote may not match the topic statement of the paragraph that it is supposed to support.

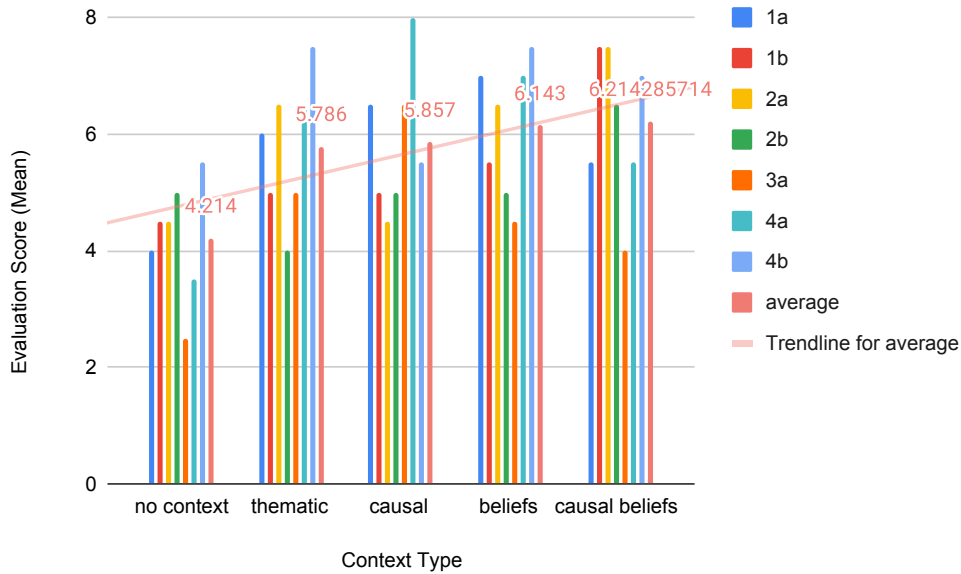


Figure 3: Mean scores for two annotators for each of the seven evaluation scenarios by context type. Five context type settings are compared. Each bar represents a mean score from the two annotators for the given scenario output produced using a given context type. The trendline is for the average of the scores for all scenarios per setting. Context setting generations outperform no context. Causal and subjective view contexts outperform the thematically-related context.

These mismatches tend to be very subtle and make the evaluation task very demanding and potentially requiring additional annotator training.

5.3.3 Annotator Differences

There was a lot of overlap in the criteria between the linguist and the domain expert, with both highlighting the use of quotations, logical flow, and organization; however, the domain expert also focused more on content (the number of factors included by the model as contributing to the explanation and the quality of the evidence provided), while the linguist gave a lot of weight to text quality and used a wide variety of text quality features as contributors to the overall score, thus lowering the weight of content quality. We view this difference in criterion prioritization from the point of view of human label variation framework discussed in (Plank, 2022). Plank views certain types of annotator disagreement as signal, for instance, when the task is subjective and open to interpretation. In our case, not only is the task complex and highly subjective, but also the two annotators come from different fields of expertise. We believe that their disagreement on the score helps us look at the performance of the system from different, complementary points of view.

5.4 Practical Constraints

With this project, we had to work within the confines of a real-world social science setting, which comes with some limitations. The first one is limited availability that domain experts face, since they may have to combine their research, teaching, and other responsibilities with on site travel for field work. With this in mind, social science experiments have to be set up to reduce the annotation load as much as possible. In our case, this meant minimizing the reading time required from each annotator. For this purpose, we set up the experiment as a spreadsheet with each scenario presented together with the five outputs instead of providing randomized scenario-output pairs, which would help avoid possible order bias (that is, the evaluator getting the impression that the stimuli are presented in a certain order, e.g., order of improved quality).

The second major limitation is the inherent subjectivity and complexity of the task. The task is cognitively demanding, with multiple competing evaluation criteria, and the length of each output.

The third limitation is the difficulty of setting up evaluation. While intrinsic evaluation that we did is possible, despite the difficulty recruiting annotators for such a cognitively demanding task, a real world evaluation is more complicated because

it may require—as it did in our case—setting up sophisticated baselines and on location field work to make sure the results are relevant for the target population. These may not always be easily accessible to computational social science practitioners, which makes the lack of extrinsic evaluation for NLG a common issue (Celikyilmaz et al., 2020; Gehrmann et al., 2023).

6 Related Work

Retrieval-augmented NLG: In exploring the integration of local information to mimic people’s behavior in query augmentation for language models, our approach is distinct from contemporary methodologies that use retrieval-augmented generation. Among these, the Corrective Retrieval Augmented Generation (CRAG) introduced by (Yan et al., 2024), employs a corrective strategy by integrating a retrieval evaluator and large-scale web searches to assess and refine the quality of retrieved documents. This method uniquely addresses the robustness of generation through corrective actions based on the quality assessment of retrieved documents, targeting the filtration of irrelevant content and enhancement of document relevance through a decompose-recompose algorithm.

In parallel, Self-RAG (Asai et al., 2024) implements a self-reflective framework that encompasses retrieval, generation, and critique. The integration of a critic model to discern the necessity of retrieval and to evaluate the quality of retrieved knowledge places emphasis on selective retrieval and the adaptive generation process. This technique contrasts with ours by focusing on the deliberation over retrieval necessity and the critique of retrieved content’s utility, aiming to optimize the generation based on retrieved knowledge relevance.

Further diverging from our approach, Aly and Vlachos (2022) and Aly et al. (2023) concentrate on the application of natural logic for enhancing reasoning in language models, particularly within question-answering contexts. Through the strategic retrieval of documents guided by natural logic, their method aims to improve logical coherence and interpretability of generated responses, marking a focus on logic-based reasoning enhancements in language generation tasks.

Contrary to the previously mentioned methods, our method involves crafting prompts enriched with diverse types of background information, encompassing subjective views of populations, causal

relations, and thematically related insights. By collaborating with linguists and domain experts, we’ve established a comprehensive set of evaluation criteria to assess the quality of these generated explanations. Our findings reveal that enriching the model’s input with targeted background information enhances the quality of its output, leading to explanations that are consistently rated higher than those generated without such contextual enrichment. This strategy not only refines the model’s ability to produce relevant and insightful explanations but also broadens the application scope of retrieval-augmented language models in understanding and elucidating complex social behaviors.

7 Conclusion

In this paper, we make predictions about and generate explanations for hypothetical socioeconomic scenarios by leveraging contextual information retrieved from the web. We show that our retrieval augmented generation approach results in outputs that are both comparable in accuracy to other LLM baselines and high in quality, as evidenced by the evaluation conducted by a domain expert and a linguist. Using subjective and causal information further improves the quality of the explanations. Moreover, by analyzing the evaluations from the two experts, we show how this is another instance of a task where differences in annotations are to be expected and encouraged thanks to its subjectivity.

8 Acknowledgments

The authors thank the anonymous reviewers for helpful discussion. This work was supported by the Defense Advanced Research Projects Agency (DARPA) under the Habitus program. Maria Alexeeva and Mihai Surdeanu declare a financial interest in lum.ai. This interest has been properly disclosed to the University of Arizona Institutional Review Committee and is managed in accordance with its conflict of interest policies. The annotation work was partially supported through Research and Project (ReaP) Grant from the University of Arizona Graduate and Professional Student Council (GPSC).

9 Limitations

There are several limitations to our work. First, we only test our approach using a single language model. With other language models, e.g., Mistral

7B, which was used for one of the baselines, the results could have been different.

Second, while English is an official language of Ghana, it is not the only widely used language in the country. By sticking to one language, we may be missing out on information that could have provided important background knowledge to the model. Moreover, we do not account for possible local variations in the use of English, except for focusing our data selection using the location-specific word *galamsey*. Used to refer to illegal, small scale mining, it has been used for information retrieval within the project since it is directly related to our use case. However, based on our analysis of the extracted causes (see Section 3.3), our NLP tools are able to extract the intended information from the retrieved data with no obvious issues.

Third, when providing background information to the model, we operate over individual sentences. By not using broader context (e.g., the full paragraph) for each sentence, we may be eliminating longer reasoning chains described in text.

Finally, our system is only marginally better than the baselines on capturing the answer distributions from local surveys. There could be multiple explanations for that. For instance, we only provide the model with a small snippet of extracted background information with each API call because of the limited token window allowable with each prompt. Larger amount of context could have resulted in better performance. People’s decisions are also not necessarily only influenced by information directly related to the question, which is what we have since we use similarity for context sentence selection: general views, e.g., people’s attitudes to the importance of legality or money in general could impact their opinions on questions about involvement in illegal mining activities. That said, we believe that providing more relevant, higher quality explanations, which our system does based on qualitative evaluation, is the main benefit of using location-specific context to prompt the model.

10 Ethical Considerations

10.1 Benchmark

The RAND benchmark model assessed how well generic LLMs could anticipate the beliefs and opinions of a local population over a set of scenario-based questions. LLMs were chosen as a comparison point to our approach for two primary rea-

sons. Firstly, they are currently de-facto automated systems for answering complex reasoning questions with minimal resources and therefore represent plausible alternatives individuals might pursue in lieu of access to the performer team model. Secondly, they are known to have limitations, as will be discussed below, that our methodology may be well-positioned to address.

LLMs are well known to replicate biases within their training sets (Feng et al., 2023; Zack et al., 2024), and may struggle to represent viewpoints of populations not well represented within them (Santurkar et al., 2023). While it is difficult to assess the degree to which Ghanans are represented in the GPT-4 and Mistral models leveraged within the benchmarks, it is safe to assume this population is represented substantially less than populations from English speaking countries. Consequently, the benchmarks scores are meant to highlight the limitations of leveraging LLMs to reflect the viewpoints of remote populations and also to illuminate how systems more rooted in data produced by local populations, such as that developed by the performer team, may address these limitations.

10.2 General Remarks

The two main concerns in regards to this work is that we are attempting to mimic responses of local populations in a different country and also that we may not be representing the views of the people of that country in a fair way. To elaborate on the second issue, it is manifested at different stages of our pipeline, e.g., we use the data sources in only one of the many local languages and from one genre (news) and we are limited by the context window size, so a lot of available information about people’s views is not passed to the model. All of this contributes to creating a potentially biased view of the population.

We attempt to minimize the bias we introduce by using local data sources and conducting evaluation by using local surveys for quantitative evaluation—even though those can also suffer from missing gender-related data, marginalizing certain groups (e.g., rural vs. urban), or biased in developing questions and translating surveys (Weber et al., 2021),—and domain experts for qualitative evaluation.

Both of these issues can be further ameliorated by involving local populations of the area investigated: both as experts to improve the quality of the tools being designed (e.g., to help identify appropriate data sources, evaluate the quality of the

outputs, etc) and as users of the tools: the tools are intended to be used by target populations to augment their decision making process and not to be used by third parties.

References

- Maria Alexeeva, Caroline Hyland, Keith Alcock, Allegra A. Beal Cohen, Hubert Kanyamahanga, Isaac Kobby Anni, and Mihai Surdeanu. 2023. [Annotating and training for population subjective views](#). In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 416–430, Toronto, Canada. Association for Computational Linguistics.
- Rami Aly, Marek Strong, and Andreas Vlachos. 2023. [Qa-natver: Question answering for natural logic-based fact verification](#).
- Rami Aly and Andreas Vlachos. 2022. [Natural logic-guided autoregressive multi-hop document retrieval for fact verification](#).
- Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. [Self-RAG: Learning to retrieve, generate, and critique through self-reflection](#). In *The Twelfth International Conference on Learning Representations*.
- Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.
- Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. *arXiv preprint arXiv:2305.08283*.
- Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Papat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.
- David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. [Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions](#). In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Barbara Plank. 2022. [The “problem” of human label variation: On ground truth in data, modeling and evaluation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cino Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.
- Rebecca Sharp, Adarsh Pyarelal, Benjamin Gyori, Keith Alcock, Egoitz Laparra, Marco A. Valenzuela-Escárcega, Ajay Nagesh, Vikas Yadav, John Bachman, Zheng Tang, Heather Lent, Fan Luo, Mithun Paul, Steven Bethard, Kobus Barnard, Clayton Morrison, and Mihai Surdeanu. 2019. [Eidos, INDRA, & delphi: From free text to executable causal models](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 42–47, Minneapolis, Minnesota. Association for Computational Linguistics.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.
- Henry Tanner. 2023. all-minilm-l6-v2. <https://github.com/henrytanner52/all-MiniLM-L6-v2>. Accessed: 2024-03-14.
- Ann M Weber, Ribhav Gupta, Safa Abdalla, Beniamino Cislighi, Valerie Meausoone, and Gary L Darmstadt. 2021. [Gender-related data missingness, imbalance and bias in global health surveys](#). *BMJ Global Health*, 6(11).
- Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*.
- Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdunour, et al. 2024. Assessing the potential of

gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6(1):e12–e22.

A Probability Calculation Walk-through Example

For example, if we have three choices that we need to rank, let's presume that we get the following rankings for the three rolls that we run: [2, 1, 3], [1, 2, 3], [3, 1, 2]. We proceed to invert the ranks so that a higher rank is better, resulting in the following ranks: [1, 2, 0], [2, 1, 0], [0, 2, 1]. The next step is to sum up the three variations into a single final ranking, resulting in the array: [3, 5, 1]. Furthermore, using $\gamma = 0.2$ and applying the $e^{\gamma \cdot \text{final_rankings}(\text{choice}_i)}$ described above we obtain the following values: [4.02, 29.68, 0.54]. The last step is to divide each value by the sum of values resulting in the final per-choice probabilities of: [11.7%, 86.7%, 11.6%].

B Survey Methodology

B.1 Data Collection Procedures

The PWLWP survey was administered by NORC's local survey firm, Ipsos Ghana, between August 16th and September 5th, 2023, via face-to-face (F2F) computer-assisted personal interviews (CAPI). The field team consisted of 17 enumerators, four supervisors, and a quality control officer who oversaw quality control activities throughout data collection. The field team was trained in-person from August 7-11, 2024 by Ipsos' field manager and trainers. NORC provided an independent consultant, who reported directly to NORC's Survey Director, to oversee training, piloting, and field launch of the survey. The survey was administered to 1,782 households in the Ashanti region of Ghana.

The English version of the survey instrument was translated using the reconciliation method (two independent translations reconciled by a third, independent translator) into Twi. Both the English and Twi versions of the survey were provided for enumerators to conduct the survey in. Ipsos Ghana enumerators, local to the region, conducted the interviews and recorded responses using tablets containing the programmed survey script in the SurveyCTO software platform. Survey data were directly uploaded from these tablets through encrypted connections to NORC's SurveyCTO cloud server on a daily basis. Data quality reviews by

NORC staff were conducted daily throughout the fieldwork period. NORC analysts shared data quality assessments with Ipsos Ghana field managers on a daily basis to allow for ongoing quality assurance and correction as needed during the fieldwork period.

B.2 Subject Population

Location: The survey was administered to randomly selected households in the Ashanti region of Ghana.

Respondents: Survey respondents included local resident adults, 18 years of age or older, who were the most knowledgeable about the household's activities in farming, animal husbandry, or mining and whose households engage in such activities on land they own, rent/lease, or borrow. The target sample size was 1,700 interviews.⁶

Inclusion and Exclusion Criteria: Anyone under 18 years of age, individuals not living in the selected household, and individuals who were not knowledgeable about the household's activities in farming, animal husbandry, or mining and whose households engage in such activities on land they own, rent/lease, or borrow were not eligible to participate.

B.3 Sampling Procedures

Sampling design: The survey used probability proportional to size (PPS) sampling, giving larger EAs in the Ashanti region a higher probability of being randomly selected for the sample. Sampling was done at the enumeration area level and not the population level, consistent with the Ghana Statistical Service sampling approach. Ipsos requested a sample frame from the Ghana Statistical Service and selected enumeration areas (EAs) using the 2021 Population and Housing Census. One hundred and seventy (170) enumeration areas (EAs) and thirty (30) replacement EA's were selected for the PWLWP survey. Ipsos obtained geo-location maps of the EAs, which provided guidance to the enumerators in locating the designated EAs and working within their defined boundaries. Prominent landmarks such as mosques, schools, markets, cattle dips, road intersections, and factories were used as reference points to mark the single starting point of the random route walk within each EA.

⁶IPSOS Ghana exceeded the target sample size and conducted 1,782 interviews in total.

Table 3 shows the breakdown of the EAs selected for the PLWPW survey.

Household and Respondent Selection: A random walk methodology was employed within each EA that ensured a random selection of households throughout the EA, to limit clustering effects. A random number (between 1 and 10) was used to start each walk path, which represents the number of households from the starting point to select first household on the path. After the initial household selection, a sample interval of 10 was used for urban areas and a sample interval of 5 was used for rural areas. Enumerators counted all houses on their left and turned around once they reached the boundaries of the EA or a dead end on their path. Once households were contacted to participate in the survey, the household was screened for eligibility. First, the respondent was asked if the household engages “farming, forestry, foraging, mining, or animal husbandry either on the household’s land or someone else’s.” If the household does, the enumerator asked to speak with “the member of the household most knowledgeable about those farming, forestry, foraging, mining, or animal husbandry activities.” Respondents were provided with a kit containing masks and soap as the incentive to participate in the survey. Each EA had a quota of 10 respondents.

Informed Consent: Prior to administering the survey, enumerators read an informed consent script to each respondent and verbal consent from the respondent was required before the survey was administered. The informed consent script included a “Right to Refuse or Withdraw” section, informing respondents that they may refuse to take part in the study at any time.

Benefits and Risks: Overall risk to respondents for participating in the survey was minimal (i.e., not greater than risks encountered in everyday life). The only known risk to the respondent was loss of time due to participating in the survey. The survey took approximately 30 minutes to administer. Respondents received no direct benefit from participation in the survey, though they were informed during the informed consent that their participation would help NORC and DARPA learn more about the perspectives of residents and activities they engage in on the lands they own, rent, or borrow in the Ashanti region in Ghana.

C Data Sources

The corpus for the experiment was compiled from the articles retrieved from the following websites:

- <https://3news.com>
- <https://www.adomonline.com>
- <https://thechronicle.com.gh>
- <https://citifmonline.com>
- <https://www.etvghana.com>
- <https://www.ghanaweb.com>
- <https://www.happyghana.com>

D RAND HABITUS Large Language Model Benchmark

D.1 Benchmark Requirements

The evaluation was created and administered by the RAND Corporation nonprofit think tank. To better contextualize performance team (PT) results for Predicting What Locals Would Predict (PWLWP) portion of the HABITUS program, RAND (TE) focused on developing a baseline model that could also be scored against NORC (CE) survey results. The main criteria we kept in mind for this baseline were:

1. Performance: We want to create a benchmark that performs similarly on PWLWP to how an uninformed operator, without access to sophisticated data sources, might
2. Simplicity/Realism: In addition to performing well, this baseline must also be relatively simple as operators generally do not currently always have access to sophisticated analysis tools or highly-curated information sources

When combined, a benchmark that satisfies both above criteria will help contextualize the value that each PT’s machinery may add to current operational challenges.

Each PWLWP scenario consisted of a multiple-choice question. Both PT and TE teams predicted the selection fraction of each response within a multiple-choice question, and these values were compared to those extracted from a CE survey of the question on a local population.

In Phase 1 of the HABITUS program, we selected a random baseline as a benchmark. For each multiple-choice question, selection fractions for each response were generated by sampling from a Dirichlet distribution, and the resulting values were compared against the CE survey results. The

Region	Total EAs	Rural EAs	Urban EAs	<i>Replacement EAs</i>	
				<i>Rural EAs</i>	<i>Urban EAs</i>
Ashanti	170	69	101	16	14
Total	170	69	101	16	14

Table 3: Enumeration Area (EA) Selection Overview for Ghana PWLWP Survey

Context Type	Generation Excerpt
1. no context	... The first part of the context notes, "China has recently completed a mining infrastructure project in Brong Ahafo, Ghana". This construction of a new mining infrastructure can bring about substantial changes in the local socio-economic structure. Inherently, such projects tend to be viewed as symbols of growth and economic opportunities, thereby attracting workers who may otherwise rely on less formal, and sometimes illegal, means of earning their livelihood. ...
2. thematically-relevant	... In addition, the Chinese government promoting the new large-scale mining effort as a source of local employment as per the context quote, "the Chinese government has promoted the new large-scale mining effort as a source of employment for locals engaged in small-scale illegal mining," could have incited locals previously involved in illegal mining to transition to the new employment opportunities. This, in turn, would account for a substantial decrease in illegal mining...
3. causal	... Another significant factor is the Ghanaian government's efforts to combat illicit mining in order to mitigate "its impact on the environment" which is "dire". With the new mining project in place and its conducive employment offers, Ghanaian miners engaged in illegal practices would now have a legal and safer alternative to contribute to the mining sector while also making a living. ...
4. belief	... The promise of formal employment in a large-scale mining operation can be seen as a powerful incentive for Ghanaians to cease their participation in illegal mining, especially since "74% of Ghanaians say no citizen should be permitted to engage in illegal small scale mining or "galamsey" for any reason". Offering an alternative, legal prospect for income can prod these individuals to abandon illegal mining...
5. causal belief	... Secondly, the Chinese galamseyers in Brong Ahafo might have willingly taken the offer to work in the new mine because of the prevalent resentment towards the Chinese for their involvement in illegal mining activities as observed in this statement: "There is already a sense of resentment for Chinese because of their involvement in illegal mining, especially in the Western Region." By transitioning to legal mining operations, they might have sought to improve their image within the local Ghanaian community. ...

Table 4: Excerpts from generated explanations for the scenario: "Imagine that China has recently completed a mining infrastructure project in Brong Ahafo, Ghana. Chinese galamseyers in the region have received an offer to work in the new mine. Additionally, imagine that the Chinese government has promoted the new large-scale mining effort as a source of employment for locals engaged in small-scale illegal mining. Suppose that within a month, the Ghanaian government has seen a decrease in illegal mining activities in Brong Ahafo. This is because..."

PT’s all showed significant improvement over this benchmark. However, while satisfying (2), the random benchmark was notably lacking in (1)—in most scenarios, an uninformed operator would be more accurate than a random guess at answering PWLWP like questions.

D.2 Language Model Benchmark

In some ways the ideal benchmark would be one that emulates an intelligent operator who responds to a PWLWP question by collecting relevant, easily-available public information and drawing a conclusion.

To simulate this process, we developed an LLM module that interfaces with Google Search to collect relevant information and produce a PWLWP response. In a slight modification to above idealized scenario, we constructed our LLM to simulate a survey fielded to residents of the PWLWP area of interest, inspired by recent work in the field (Argyle et al., 2023).

The LLM is told to consider the identity of a resident of the region and then asked to express their beliefs on the PWLWP question, and this process is repeated to mimic the results of a random survey of the population. The LLM module is fairly bare-bones and leverages off the shelf programming open-source Python libraries with minimal custom coding. Pseudocode is provided in Table 5.

Much like in human surveys, we find randomizing the ordering of the response options for each LLM call prevents biases towards picking the first or last option. Descriptions of the main components of the LLM module are provided in Figure 4.

Language Model: We conducted our benchmark evaluations using the GPT-4 library. We interface with the model through the popular open-source library LangChain⁷.

Google-Search: We equip the LLM with the ability to conduct Google Searches to retrieve information needed to better contextualize the question that lies outside of the GPT-4 training set. The LLM is providing access to Google through the Agents module within LangChain. This module prompts LLM to devise a strategy to solve a given problem and then provides the LLM with tools, such as Google Search to execute its strategy. Each time the LLM collect new information from its

tools, it updates its understanding of the problem, revises its strategy if needed, and proceeds until a desired result is achieved.

Response Ranking: Rather than merely selecting the most likely option amongst a set of multiple-choice responses, we ask the LLM to rank the responses in order of decreasing likelihood. This allows us to better understand LLM decision making and accordingly make prompt adjustments. Also, as will be discussed below, this may play a role in producing distributions of selection fractions across all response options.

Chain of Thought Prompting: We leverage the popular technique of chain-of-thought prompting within our LLM calls. This technique instructs LLMs to provide reasoning for decision making, and in certain cases, has been found to improve LLM performance while providing information relevant for debugging/prompt engineering.

Identity Contexts: For each PWLWP question, we conduct $N \sim 25$ simulated surveys. Each survey instructs GPT to assume the identity of a citizen of a given region with certain demographic characteristics. For example, if a PWLWP question was fielded to residents of Ghana, for each LLM call, we provide the age, gender, and city of residence of a ‘synthetic citizen’ of Ghana. In this case, these demographic variables are sampled from distributions (1) publicly available online or (2) provided in advance by the CE team.

Admittedly, it may be difficult to extract accurate demographic information for certain areas of interest, and further there may be complex relationships between demographic variables that need to be considered when sampling. Both factors pose challenges to exporting this technique generally. However, this component is not essential to the LLM module and further experiments can investigate how the performance of the benchmark changes, if at all, if identity contexts are excluded.

D.3 Prompt Design

An approximate form of the prompt we developed is provided in Figure 5.

The Google Search tooling, chain of thought prompting, response rankings, and identity contexts mentioned above are all visible from within the prompt.

⁷<https://www.langchain.com/>

LLM Module
For k in N:
<ul style="list-style-type: none"> • Randomly sample an identity context I • Randomize the ordering of the response options R • Process LLM(P, R, I) and store result
Aggregate the N results from above into a final selection fraction distribution

Table 5: Pseudocode for the LLM Module. N is the number of LLM samples, LLM(P, R, I) is the result of the LLM module called on PWLWP question P with given response options R conditioned on identity context I.

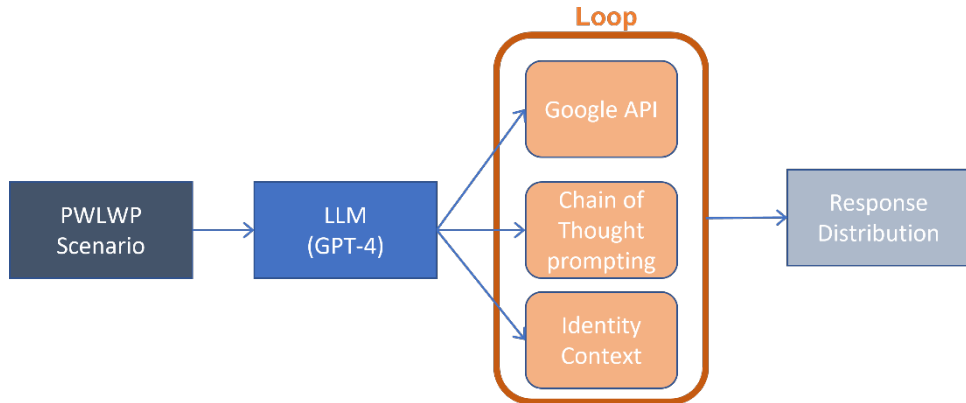


Figure 4: The overview of the benchmark LLM module

D.4 Distribution Generation

One challenge we encountered was extracting selection fraction distribution from our LLM module. If there are four answers to a given PWLWP question: A, B, C, D – we need to predict the fraction of individuals who will select each response. One could ask the LLM to produce these anticipated fractions as a response, yet LLMs have struggled to produce well-calibrated probabilities in certain scenarios (Srivastava et al., 2022). We present two alternative options here:

Top Vote Aggregation (‘GPT-4 TopVote’):

Across our N LLM samples per PWLWP question, we track how often each multiple-choice option gets selected as the top choice. The final selection fractions are the percentage of time each response gets selected as the ‘top vote’. One question that emerges from this technique is “how many votes is enough?”. While we can monitor the convergence of each selection fraction as a function of LLM call, it is still difficult to anticipate how these fractions may change under additional samples. Further, the GPT-4 calls used as a foundation for our model are both relatively slow and expensive (~ 1 min per sample, about $\sim \$0.50$ dollar per sample), meaning dramatically increasing the number of samples may

pose challenges.

Ranking Calibration (‘GPT-4 Calibrated’):

For each vote, we may instead assign a weighting to each response based on the probability ranking of that response within the LLM output. In this case, the calibration selection fraction, SF , for response r within a PWLWP question is given as:

$$SF(r) = \frac{\sum_i^N w(\alpha_{r,i})}{\sum_r^R \sum_i^N w(\alpha_{r,i})}$$

where $\alpha_{r,i}$ is the ranking of response r on the i^{th} LLM vote, w is the weighting function that assigns weights based on $\alpha_{r,i}$, and R is the entire response set for a given PWLWP question. In this work, we chose

$$w(\alpha) = Ae^{-\alpha \cdot b}$$

where A and b are constant adjusted to our data. In an ideal scenario, other functional forms would be explored and A and b would be fit to a more robust calibration dataset (here a ranking of 0 “top-vote” get assigned the highest weight, with decreasing rates with increasing rankings).

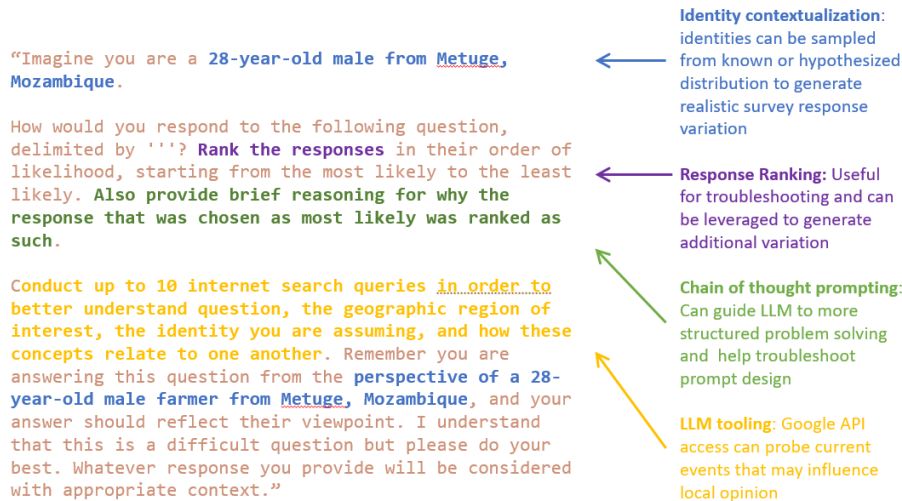


Figure 5: A sample prompt for the LLM benchmark for a sample location.

D.5 Analysis of PWLWP Results

We deployed our LLM benchmark on HABITUS Phase II PWLWP questions fielded to the region of Ghana and compared our results against both PT and CE teams. There were seven total scenarios in this set, with one multiple choice question each. The ultimate goal is to determine how similar the PT and TE results are to the CE results across these questions. However, similarity can be defined in multiple ways. For this evaluation, we used mean absolute error (MAE) as described below:

$$MAE = \frac{1}{|R|} \sum_r^R |sf_{r,CE} - sf_{r,K}|$$

where $|R|$ is the number of responses in set R and $sf_{r,K}$ is the selection fraction associated with response r produced by team K (here $K = CE, PT, \text{ or } TE$).

D.6 Offline Model Benchmarks

The previous LLM benchmark utilized closed-source GPT models that are only accessible via an online API. Eventual end users of HABITUS technologies may need to access model results in edge computing environment without access to Google searches/online APIs and may also need to transmit classified information to such technologies that cannot be shared to third party vendors.

To provide a benchmark that satisfies these operational requirements, we built a second LLM pipeline that leverages only open-source models that can be run locally and offline. Our benchmark utilized the 7B parameter Mistral AI language model.

We evaluated PT PWLWP scenarios using essentially the same setup as above with this Mistral model, with two main differences. Firstly, we removed the ability for the model to conduct Google searches. Secondly, we removed the response ranking component of the model, asking instead that the model simply return the option it deems most probable during every LLM vote. Part of the motivation of the response ranking was to generate variation in the LLM output given that GPT-4 model calls can have high latency and high costs compared to open-source models. With Mistral, we can conduct 100 model votes per scenario quite easily, allowing us to generate response variation more naturally without adding in additional post model calibration. Lastly, the response ranking adds additional complexity to both the LLM prompt and the structure of the LLM output, complexities that may pose challenges for the open-source model. By removing these components, we make our pipeline much simpler and reduce the possibility of error.

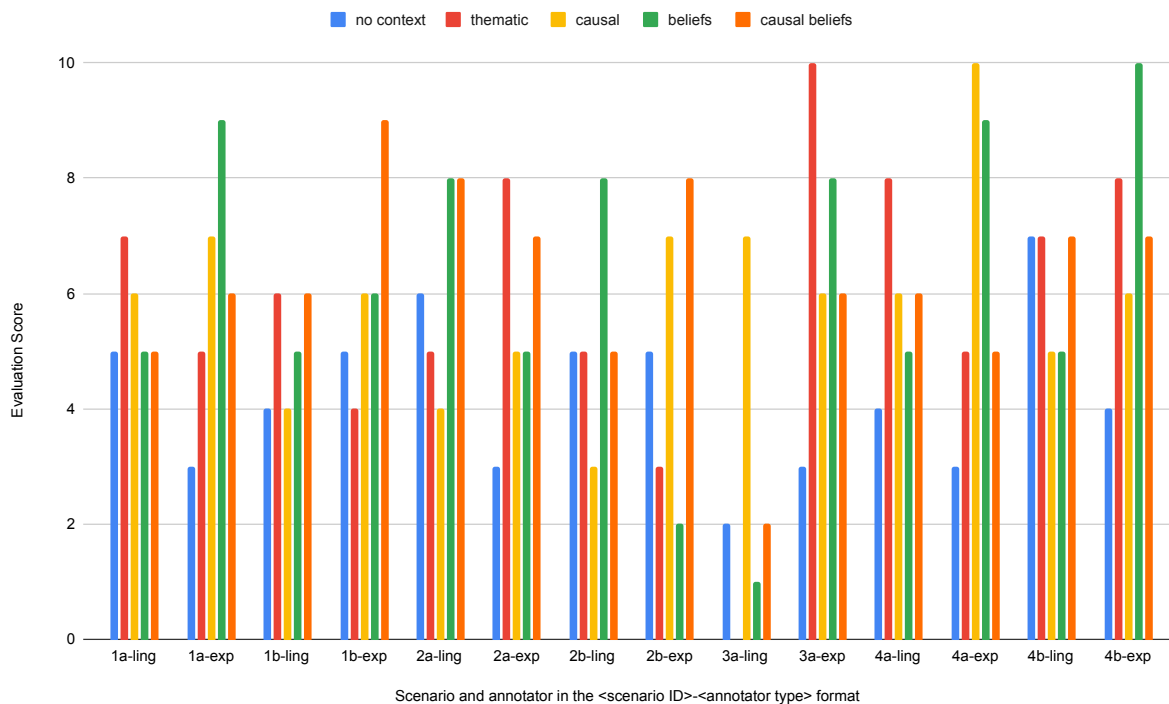


Figure 6: Breakdown of scores by two annotators per scenario. In the scenario and annotator code, “-exp” stands for domain expert and “ling” for linguist. In the corresponding scenarios (e.g., the two scenarios starting with 1a), the corresponding generated explanations are in the same color and order. For instance, in scenario 1a, the linguist gave the no context setting (blue color, first bar) the score 5 and the expert the score 3. We attribute the difference in scores to the two annotators prioritizing different types of criteria.

Context Type	Scenario (paraphrase)	Generation Excerpt
1. subjective view context	Illegal mining causing land degradation and shift of labor away from agriculture resulted in mass protests and tension. Explain how that impacts illegal mining.	...Second, illegal mining activities can cause severe environmental damage, leading to the land’s degradation and ruining the area’s potential for future agricultural activities. According to the context, "it is acknowledged that while illegal mining supports livelihoods of many, it has equally caused severe damages to the environment in many parts of the district." Moreover, ...
2. causal context	Explain lack of decrease of illegal mining in a given location after a foreign country builds infrastructure for and promotes legal mining.	...Second, there is widespread corruption and ongoing policy implementation challenges. The context reports "widespread corruption has hamstrung government efforts" in fighting against illegal mining, referring to illegal mining as a "menace" that may deter engagement in legal opportunities. ...

Table 6: Examples for Section 5.3. Ex. 1: The quote is relevant for the topic sentence of the paragraph, but is only connected with a transition phrase and not elaborated on; additionally, the explanation does not match the scenario. Ex. 2: Use of short snippets of quotes from provided context, which, while successful, is harder to verify.

Where on Earth Do Users Say They Are?: Geo-Entity Linking for Noisy Multilingual User Input

Tessa Masis (*they/them*) Brendan O’Connor
University of Massachusetts Amherst, MA, USA
{tmasis, brenocon}@cs.umass.edu

Abstract

Geo-entity linking is the task of linking a location mention to the real-world geographic location. In this paper we explore the challenging task of geo-entity linking for noisy, multilingual social media data. There are few open-source multilingual geo-entity linking tools available and existing ones are often rule-based, which break easily in social media settings, or LLM-based, which are too expensive for large-scale datasets. We present a method which represents real-world locations as averaged embeddings from labeled user-input location names and allows for selective prediction via an interpretable confidence score. We show that our approach improves geo-entity linking on a global and multilingual social media dataset, and discuss progress and problems with evaluating at different geographic granularities.

1 Introduction

The real-world geographic location of social media users is valuable data for many computational social science tasks, including disaster response (Kumar and Singh, 2019), disease surveillance (Lee et al., 2013), analyzing language variation (Huang et al., 2016), and comparing regional attitudes (Rosenbusch et al., 2020). Many studies have used Twitter (now known as X) data for such analyses, focusing on geo-tagged tweets where each tweet is associated with latitude and longitude coordinates. However, geo-tagging with coordinates was deprecated in June 2019 and even before then only a small percentage of tweets (< 2%) was geo-tagged (Kruspe et al., 2021).

It has thus become increasingly necessary to infer location from user profiles and especially from the free text Location field, in which a user may enter anything they want to identify their location. This field is frequently specified, with at least 40% of users providing recognizable locations in over 60 different languages (Huang and Carley, 2019b).

The task of linking a location reference to the actual geographic location is known as *geo-entity linking* (see Table 1 for examples). There are few open-source multilingual geo-entity linking tools available and existing ones are often rule-based (Alex et al., 2016; Dredze et al., 2013), which may break easily in noisy social media settings, or LLM-based (Zhang et al., 2023), which are too expensive for large-scale datasets.

In this paper, we investigate the task of geo-entity linking for noisy, multilingual user-input location references. Our work makes the following contributions:

- We propose a method for geo-entity linking of noisy and multilingual user input by representing real-world locations with averaged embeddings from labeled user-input location names. Unlike previous methods, ours enables selective prediction via an adjustable threshold for cosine similarity scores, which we analogize with confidence scores (§4).
- We compare performance of multiple variations of our proposed method on a global and multilingual dataset, and show that all of them outperform the leading baseline (§5).
- Through a manual annotation experiment, we approximate accuracy upper bounds on our dataset and show that our method is near the upper bound at country- and administrative-levels but quite far below at the city-level. We discuss problems with geo-entity linking social media data at the city level (§6).

2 Related Work

Geo-entity linking, also known as toponym resolution, seeks to link some mention of a geographic entity to the correct entity in a target database. Previous approaches typically use some combination of the mention’s text and/or context; knowledge

User-input location	Real-world location	Type of noise
TURKEY/SINOP	Sinop, Sinop, TR	Uncommon punctuation use
福島県いわき市	Iwaki, Fukushima, JP	Non-Latin script
Catskills	Hyde Park, New York, US	Informal/alternative name
where the wild things are	N/A	Not a real location

Table 1: Examples of user-input location references, the real-world locations they should be linked with, and the type of noise that the geo-entity linking model must be able to handle.

bases (e.g. gazetteer, Wikipedia) which contain features such as population, location type, etc.; and coordinates/geometric features. They may use some mix of rule-based, unsupervised, and supervised methods. The majority of prior work on geo-entity linking has focused on data in the English language and in the domain of news articles (Lieberman and Samet, 2012; Speriosu and Baldrige, 2013; Kamaloo and Rafiei, 2018; Cardoso et al., 2019; Kulkarni et al., 2021; Cardoso et al., 2022; Sá et al., 2022; Li et al., 2022; Sharma et al., 2023; Zhang and Bethard, 2023). We note that the geospatially grounded model GeoLM – which was trained only on English data – was evaluated on geo-entity linking in a way similar to our proposed method, by ranking locations by cosine similarity between each candidate and the query location (Li et al., 2023). This is similar to our proposed NameGeo method, although the authors did not explore using cosine similarity thresholds for selective prediction or any of the other variations that we investigate.

Some prior work has examined geo-entity linking in historical texts (Smith and Crane, 2001; Ardanuy and Sporleder, 2017; Ardanuy et al., 2020), which includes English, Spanish, Dutch, and German data; and in web pages (Moncla et al., 2014), which includes French, Spanish, and Italian data.

Most relevant to the current work are previous studies which have examined geo-entity linking in social media data. Alex et al. (2016) uses a rule-based English system, Dredze et al. (2013) uses a multilingual rule-based system, and Zhang et al. (2023) uses a multilingual LLM-based system.

Entity-linking is the broader task of linking some mention of an entity – which could be a person, place, or organization – to the correct entity in a target database. We note a similarity between our proposed UserGeo method and one introduced for entity-linking in FitzGerald et al. (2021), in that they both represent entities using all mentions in the training data. However, our method is simpler as it does not involve model training. Addition-

ally, UserGeo represents an entity by averaging mention embeddings instead of having a separate embedding for each mention, which induces a more holistic entity representation that can better handle noisy mentions present in social media data.

Predicting user location, also known as user geolocation, is a task distinct from geo-entity linking social media data in that it seeks to determine the location of a user using both text data and user metadata, including post content, user bio, user language, time zone, or social networks (Han et al., 2012; Jurgens et al., 2015; Rahimi et al., 2015; Huang et al., 2016; Rahimi et al., 2017; Izbicki et al., 2019; Huang and Carley, 2019a; Luo et al., 2020). The Location field is one of many features used for location prediction, if used at all.

Geoparsing is the task of both identifying and linking geographic entities in unstructured text (Wang and Hu, 2019), essentially combining toponym recognition (Hu et al., 2023) and geo-entity linking. Geoparsers focused on standard English texts (including news articles, Wikipedia, or scientific papers) include CLIFF-CLAVIN (D’Ignazio et al., 2014), TopoCluster (DeLozier et al., 2015), CamCoder (Gritta et al., 2018), and DM_NLP (Wang et al., 2019). Other geoparsers include the Edinburgh geoparser for historical English text (Grover et al., 2010), GeoTxt for English social media data (Karimzadeh et al., 2013, 2019), and Perdido for French texts (Moncla et al., 2014).

3 Task and Data

3.1 Geo-entity linking task

Given a target location database D , a training set T containing user-input location name and ground truth location pairs, and a test set of user-input location names I , for each $i \in I$ we model

$$\arg \max_{d \in D} \text{score}(d, T, i)$$

to predict best matching geographic entity d ,

which is represented by a triple containing a city name, primary administrative region name (e.g. state, province), and country name. (The city and administrative region names may be empty strings if the entity is of a higher granularity, e.g. a country) The `score()` function evaluates the quality of d as a match to user input i , given training data T .

We note that a user-input location name i may contain multiple locations or no real locations. Predicted entity triples are allowed to be composed of only empty strings (referred to as NULL), indicating that no location could be predicted for the given user input.

3.2 Data

Target location database. We use a modified version of the GeoNames¹ database, which contains entity names and coordinates for over 11M countries, administrative regions, counties, and cities across the globe. We filtered this database to exclude cities with populations under 15K, since tweets are more likely to come from more populated areas. Our final target location database contains 28,767 distinct locations: 252 countries, 3,947 administrative regions, and 24,568 cities.

Train and test dataset. We use data from the TWITTER-GLOBAL dataset (Zhang et al., 2022). The original dataset is described as containing data from 15.3M tweets which are either tagged with geocoordinates or Twitter Place objects and are posted by users with a non-empty Location field. The tweets were posted from 2013 to 2021 and contain global and multilingual data.

We use only the 4.1M geocoordinate-tagged tweets in TWITTER-GLOBAL, because geocoordinates are much more reliable than Place objects. Geocoordinates are meant to be the exact geolocation of the user’s device and always specify a precise latitude and longitude, while Place objects are pre-defined geographic entities that the user selects from a list and may be at any granularity, from points of interest to countries (Kruspe et al., 2021). Users may assign a Place to a post simply because they are talking about it and not because they are actually there. To identify the ground truth location for each geocoordinate-tagged tweet, we used the *reverse-geocoder* library to map the coordinates to the closest city in our target location database.² There are tweets from 196 different

¹<https://www.geonames.org/>

²<https://github.com/thampiman/reverse-geocoder>

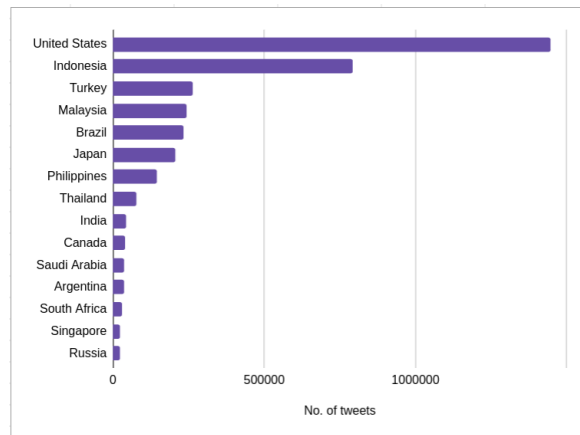


Figure 1: Tweet frequency for top 15 countries in the geocoordinate-tagged subset of TWITTER-GLOBAL.

countries, although they are not evenly distributed (see Fig. 1 for tweet distribution of the 15 most frequent countries).

We note limitations to this data. Geo-tagged tweets are not a random or representative sample of all tweets and there have been shown to be biases in who uses geo-tagging. Huang and Carley (2019b) show that less than 3% of Korean-speaking users geo-tag their posts while more than 40% of Indonesian-speaking users do. Certain countries like Turkey and Indonesia have very high percentages of coordinate-tagged tweets out of all geo-tagged tweets (53% and 67%, respectively), where most countries have 15-30%. In addition, users who have non-empty Location fields are more likely to use geo-tags, and there is strong homophily in geo-tagging behavior where users tend to connect to friends with similar geo-tagging preferences. Pavalanathan and Eisenstein (2015) show that users who geocoordinate-tag their posts versus have non-empty Location fields are measurably different groups in terms of demographics like age and gender.

It has also been shown that a non-empty Location field does not always correspond to the geo-tagged coordinates. Alex et al. (2016) find that in their dataset, 40% of users have geo-tagged coordinates within 10km of their specific Location, 70% are within 100km, and 85% are within 1000km. Despite these limitations, previous work studying geo-entity linking and user geolocation has found geocoordinate-tagged data to be useful, so we use such data here.

4 Methods

4.1 Proposed Method

We propose a method (referred to as UserGeo) that computes embeddings for each location in the target database and then, to predict a location for some user input, simply predicts the location with the closest embedding. To compute the target location embeddings, UserGeo uses training set T which contains a pair (x_t, y_t) for each tweet t , where x is the user-input Location field and y is the ground truth location triple. To help supplement locations with few user inputs, we additionally create a pair $(s(d), d)$ for each $d \in D$ where $s(d)$ is a string representation of d , with comma-separated city, primary administrative region, secondary administrative region, country name, and two-letter country code, as applicable. UserGeo creates embedding representation Z_d for each location $d \in D$ by averaging all associated x_t and $s(d)$ embeddings:

$$Z_d = \frac{1}{|\{t : y = d\}|} \sum_{t:y=d} e(x_t)$$

where $e(x)$ is the embedding for x from a pre-trained language model.

Then, for each user-input Location field $i \in I$, its predicted location triple \hat{d}_i is the location embedding Z_d that it has the highest cosine similarity with. If the cosine similarity with all location embeddings is below a given threshold t , then this is interpreted as a low confidence score and no prediction is made. In other words,

$$m = \max_{d \in D} c(Z_d, e(i))$$
$$\hat{d}_i = \begin{cases} \arg \max_{d \in D} c(Z_d, e(i)) & \text{if } m \geq t \\ \text{NULL} & \text{otherwise} \end{cases}$$

where $c(a, b)$ is the cosine similarity between vectors a and b .

The motivation behind this method is that it leverages millions of examples of user-defined location names, essentially inducing a soft-alias location name database.³ We note parallels between our framework and two other methodological classes. First, Bi-encoders obtain separate embeddings for

³We don't investigate fine-tuned models, due to two main disadvantages. First, it is unclear exactly what the fine-tuned model would learn (e.g. learning regional slang, instead of variable ways of expressing a location). Second, it is more likely for the model to overfit to the training data and perform poorly on countries that are not well-represented in it.

two sentences and then calculate cosine similarity between them. Second, Gaussian Discriminant Analysis classifies a point based on the minimum distance to clusters learned from training data.

4.2 Baselines

We evaluate four variations of our method as baselines. NameGeo: for each location $d \in D$,

$$Z_d = e(s(d))$$

. In other words, a zero-shot version of UserGeo where each target location is represented only by its embedded string name.

Different embedding models: we evaluate embedding with three SBERT (Reimers and Gurevych, 2019) variants – the popular *all-MiniLM-L6-v2* model, the multilingually-trained *paraphrase-multilingual-miniLM-L12-v2* model, and the larger *all-mpnet-base-v2* model – and the geospatially grounded GeoLM (Li et al., 2023).

Variants: there are multiple $s(d)$ functions to create different string representations of d that are all included in Z_d . The original $s(d)$ represents d as only "<city>, <admin2>, <admin1>, <country>, <2-letter country code>". Here, if d is a country then it is also represented as "<country>"; if d is a primary administrative region, then it is also represented as "<admin1>", "<admin1> in <country>", and as "<country> / <admin1>"; and if d is a city, then it is also represented as "<city>", "<city> in <admin2> in <admin1> in <country>", "<admin1> / <city>", and as "<country> / <city>".

Pruning: removing outlier user inputs x_t from each Z_d . A user input is determined to be an outlier if the squared Euclidean distance between $e(x_t)$ and Z_d is farther than a given threshold.

We also evaluate the only prior open-source tool that was explicitly created for and evaluated on broadly multilingual data. Carmen 2.0: uses a combination of regular expressions and manually curated aliases to predict real-world locations (Zhang et al., 2022).

5 Experiments

5.1 Experimental setup

Data. We divide the Twitter-Global data into a 90/10 split, with 3.7M examples in the training set and .4M in the test set. We evaluate at three levels of geographic granularity (city, primary administrative region, and country). A predicted triple is correct at the country level if the predicted country

Method	Country	Admin.	City
CARMEN 2.0	43.5	27.3	9.8
<i>all-MiniLM-L6-v2</i>			
NameGeo	59.8	37.7	14.3
UserGeo	67.8	44.2	14.8
<i>all-mpnet-base-v2</i>			
NameGeo	60.9	38.3	14.9
UserGeo	<u>67.4</u>	<u>43.7</u>	13.9
<i>paraphrase-multilingual-MiniLM-L12-v2</i>			
NameGeo	48.7	28.9	8.1
UserGeo	57.0	34.3	9.4
GEOLM			
NameGeo	52.5	30.5	12.1
UserGeo	57.4	33.9	10.7

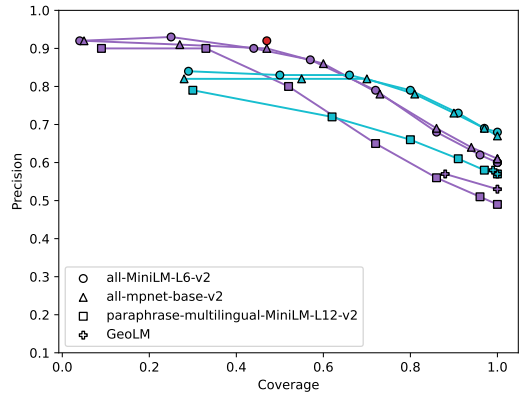
Table 2: Accuracy scores (%) at all three geographic granularities for Carmen 2.0, and NameGeo and UserGeo with different embedding models (all at threshold of 0, i.e. predictions are made for all examples). The highest score is bolded, the top 2 are underlined.

is a string match with the correct country; correct at the administrative level if both predicted country and administrative region are string matches with the correct ones; and correct at the city level if predicted country, administrative region, and city are all string matches with the correct ones. This hierarchical string matching procedure has limitations in that it does not remove all geographic ambiguity (e.g. if there are multiple cities with the same name in the same administrative region), but it should be effective in the vast majority of cases.

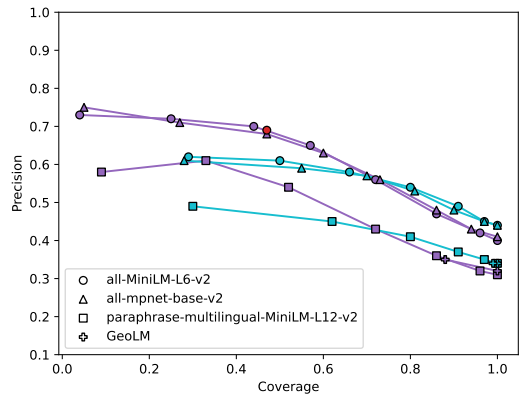
Metrics. We use four metrics to evaluate methods. *Accuracy* is the percentage of examples for which the method made a correct location prediction. Since many models do not always predict a location for all user inputs, we also evaluate *coverage* (the percentage of examples with a non-NULL prediction) and *precision* (the percentage of correct non-NULL predictions, following the standard definition of precision). For individual country performance we also evaluate F1-score, since most countries are uncommon.

5.2 Results

NameGeo, UserGeo, and Carmen 2.0. UserGeo achieves the highest accuracy at the country and administrative level, with gains over Carmen 2.0 of 25 and 17 points, respectively, and NameGeo achieves the highest accuracy at the city level, with gains over Carmen 2.0 of 5 points (Table 2). And while Carmen 2.0 has a competitive precision-coverage



(a) Country



(b) Admin.

Figure 2: Precision-coverage curves at the country (a) and administrative (b) levels. Red points are Carmen 2.0, purple are NameGeo, and cyan are UserGeo. NameGeo and UserGeo are shown with four different embedding models, where each point represents the precision and coverage at a threshold $t \in \{0, .1, .2, .3, .4, .5, .6, .7, .8, .9\}$. These curves demonstrate how thresholds can be used to tradeoff between precision and coverage with NameGeo and UserGeo.

tradeoff, especially at the country and administrative level, its overall coverage is still quite low (see Fig. 2, and Fig. 4 in Appendix). In other words, it is often correct when it makes a prediction but it does not often make a prediction. It also only has a single precision-coverage point (similar to most other prior tools). In contrast, our proposed methods demonstrate the ability to tradeoff between precision or coverage by choosing a different threshold and can achieve a higher precision or coverage than Carmen 2.0 at certain thresholds. The threshold can be used to adjust the precision/coverage balance if, for a given application, it is more important to get predictions correct or if it is more important

User-input	Carmen 2.0	NameGeo @0.5	UserGeo @0.5
TURKEY/SINOP	NULL	"" , Sinop, TR	Boyabat, Sinop, TR
福島県いわき市	NULL	Zhongshu, Yunnan, CN	Iwaki, Fukushima, JP
Catskills	NULL	Catalca, Istanbul, TR	Greenburgh, New York, US
where the wild things are	NULL	NULL	NULL

Table 3: Error analysis of the same user-input examples as in Table 1 (see Table 1 for corresponding real-world locations). Results from NameGeo and UserGeo are using the *all-MiniLM-L6-v2* SBERT model. Empty strings indicate that the model did not make a prediction at that geographic granularity, and NULL indicates that no prediction was made.

to make more predictions.⁴

We conducted a manual error analysis comparing Carmen 2.0, NameGeo, and UserGeo and observed trends in the types of errors made by each method (see Table 3). First, Carmen 2.0 rarely makes predictions for user-inputs with unexpected punctuation or in non-Latin scripts. Second, NameGeo often incorrectly predicts locations that look superficially similar to the user input (e.g. it predicts a location in China for a user input written in Japanese, and a location named ‘Catalca’ for the user-input ‘Catskills’). Third, UserGeo often correctly predicts locations for non-Latin inputs and alternate/informal location names. And lastly, all three models are frequently able to identify user-inputs that are not real locations.

We compare country-level F1-scores⁵ across countries for Carmen 2.0, NameGeo, and UserGeo, for the 23 countries with more than 1,000 examples in our test set, in order to investigate geographic bias in our models (Liu et al., 2022). We observe that the number of examples per country in the training set – which may differ by multiple orders of magnitude – does not appear to influence UserGeo performance (see Table 6 in Appendix). This suggests that an unbalanced training set doesn’t negatively impact performance as it might for a traditional supervised learning method.

Different embedding models. Across SBERT bases, we find that the *all-MiniLM-L6-v2* model surprisingly performs better than the multilingually-trained *paraphrase-multilingual-MiniLM-L12-v2* model and it performs comparably with the larger *all-mpnet-base-v2* model (Table 2, Fig. 2). Anecdotally, we found that the multilingual model per-

⁴We note that these precision-coverage curves are valid for only this dataset. Users would likely have to reevaluate precision and coverage on a new dataset or domain in order to choose an appropriate threshold.

⁵We use F1-score instead of accuracy here in order to better represent rare countries.

forms worse for examples in Latin script; for example, NameGeo with the multilingually-trained model incorrectly predicts ("" , Pukapuka, CK) for user input "Kucukyali izmir" while NameGeo with *all-MiniLM-L6-v2* correctly predicts (Izmir, Izmir, TR). On the other hand, it performs better for non-Latin scripts; UserGeo with the multilingually-trained model correctly predicts (Dnipro, Dnipropetrovsk Oblast, UA) for "Днепрпетровск" and partially correctly predicts (Varto, Mus, TR) for "إسطنبول - تركيا", while UserGeo with *all-MiniLM-L6-v2* only partially correctly predicts (Pidhorodne, Dnipropetrovsk Oblast, UA) and incorrectly predicts (Tabuk, Tabuk Region, SA). We hypothesize that part of why the improved performance for non-Latin scripts does not outweigh the decrease in performance for Latin script is that non-Latin user inputs are a minority in this dataset; for a different dataset, it’s possible that the performance of the multilingually-trained model may be better than that of *all-MiniLM-L6-v2*.

Regarding the GeoLM model, the only SBERT model it outperforms is the multilingual one. Additionally, the cosine similarity threshold does not work effectively for GeoLM because the cosine similarities between a user input and each target location are very close; for example, the average cosine similarity between a correct prediction and an incorrect prediction for NameGeo with GeoLM at the country-level is .95 and .92, respectively. There is therefore a very limited range of values for which to have a threshold. Fig. 2 demonstrates this limited range with very short precision-coverage curves for GeoLM.

Since the *all-MiniLM-L6-v2* model performs the best out of all the embedding models, we use it for the rest of the experimental results.

Variants. Adding additional variants of the location name improves performance for NameGeo,

and in fact NameGeo+variants does better than UserGeo at the city-level (see Table 4). However, adding additional variants does not improve performance for UserGeo, and UserGeo+variants does marginally worse than UserGeo.

We investigate the impact of number of location mentions in the training data on model performance. Fig. 3 contains a plot of the number of mentions in the training data per location versus NameGeo/UserGeo accuracy. NameGeo does not use any user inputs from the training data so it can be considered a control for how training data affects performance in UserGeo. We see that while some training data is better than none (UserGeo consistently outperforms NameGeo when number of mentions is less than 1,000), it is not true that more training data always continues to improve performance (UserGeo and NameGeo have comparable performance when number of mentions is greater than 1,000). This suggests that having more variants or examples in a location’s averaged embedding is helpful, but only to an extent. Few-shot is better than zero-shot, but after a certain point there are diminishing returns.

Pruning. We defined an outlier as an embedding that was more than the average distance from its centroid, where the average is calculated using all embeddings associated with the given location. On average, 38% and 42% of embeddings in each cluster were pruned for UserGeo and UserGeo+variants, respectively.⁶

In general, removing outliers did not improve performance and frequently made it worse (Table 4). This suggests that having more variety in the averaged embeddings is good for performance, even if it comes at the cost of noise.

We observed via error analysis that it is often the case that user inputs close to the centroid will be standard English spellings of the location, ones within 1x average distance will include the location spelled in different scripts or nearby/related locations, and ones more than 1x average distance will not be very related to the location. For example, for the country Armenia, user inputs less than .5x average distance include "Yerevan, Armenia" and "ARMENIA"; user inputs less than 1x average distance include "Armenia | Հայաստան | Армения"

⁶We determined through preliminary analyses that the average distance was an appropriate threshold, e.g. with a threshold of 2x the average distance, only 3% of embeddings in each cluster were pruned on average and there was no improvement in performance.

Method	Country	Admin.	City
NameGeo	59.8	37.7	14.3
+variants	62.0	40.9	17.0
UserGeo	67.8	44.2	14.8
+pruning	63.5	41.4	13.2
+variants	<u>66.0</u>	<u>43.7</u>	<u>15.3</u>
+variants+pruning	65.2	43.4	13.9

Table 4: Accuracy scores (%) for variations of NameGeo and UserGeo (all at threshold of 0); the highest score is bolded, the top 2 are underlined. The *all-MiniLM-L6-v2* SBERT model was used for all results.

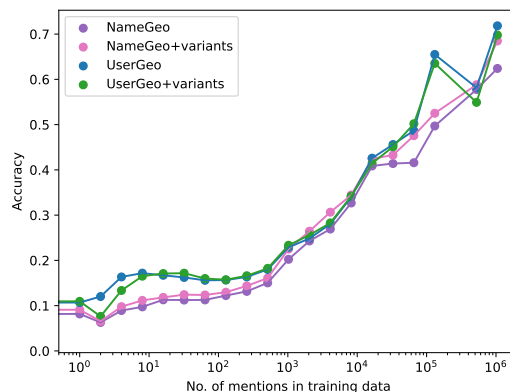


Figure 3: Average accuracy for a given number of mentions in training data (bucketed by $\lfloor \log_2 \# \text{mentions} \rfloor$), for NameGeo and UserGeo with and without adding location name variants. Location entities at all 3 geographic granularities are present in the plot.

and "Azerbaijan, Baku"; and user inputs more than 1x average distance include "Paris/Singapore" and "Worldwide".

However, it is also frequently the case that user inputs one may want to exclude are less than 1x average distance and inputs one may want to include are more than 1x average distance. For example, "Moscow" is less than 1x average distance and "ARMENI, ABOBYAN." is more than 1x average distance. User inputs that are semantically relevant (i.e. refer to the location) but stylistically dissimilar (e.g. contain emojis or uncommon punctuation uses) are often farther from the centroid than user inputs that are less semantically relevant but more stylistically similar (e.g. "New York" is within 1x average distance of the centroid for the country Gibraltar, while "G I B R A L T A R" is not). These results suggest that the model is not able to effectively differentiate between good noise and bad noise, and thus it is better to not do pruning

and to keep as many user inputs in the training data as possible.

6 Accuracy Upper Bounds

To approximate an upper bound for accuracy scores on our dataset, we conducted an analysis of a random sample of user inputs to determine how many contained references to actual locations. The percentage of examples which contain a location at each granularity suggest an upper bound for accuracy, indicating the approximate proportion that a geo-entity linker could be expected to connect to a real geographic location. The first author manually annotated 120 random examples from the training set for whether or not the user input contained an identifiable reference to a location at the country, administrative, and/or city levels.⁷

We include results from this annotation in Table 5 as well as a comparison with our best performing models’ accuracy scores. UserGeo is only 5 points below the country-level upper bound and 14 points below the administrative-level upper bound, indicating that our current performance is fairly close to the upper limit in terms of accuracy. In contrast, NameGeo+variants is more than 30 points below the city-level upper bound.

We discuss a few problems with evaluating geo-entity linking at the city level for social media data. First, the assumption of the geocoordinates as ground truth is frequently untrue and especially so at the fine-grained city level. Unlike news articles or Wikipedia data, the place that a user puts on social media is not necessarily the place where they actually are. As Alex et al. (2016) have shown, only 70% of users have geocoordinates within 100km of the place specified in their Location field. Second, since we define ground truth as the closest city to the geocoordinates with a population higher than 15K, there will always be a mismatch between ground truth and Location field if a user puts a city with a lower population. Third, there is a disconnect between what is considered a city by a database such as GeoNames versus by an everyday person. Istanbul, Jakarta, Moscow, Gaborone, Lampung, and Santa Ana are all names of both cities and administrative regions in GeoNames – when users put them in their Location field, it’s unclear

⁷We did not count coordinates as containing a location, as we wanted to investigate the ability of the model to identify natural language location references. If a location was named that could be either an administrative region or a city, then the example was marked as containing both.

	Country	Admin.	City
Upper bound	72.5	58.3	49.2
NameGeo+variants	62.0	40.9	17.0
UserGeo	67.8	44.2	14.8

Table 5: Accuracy upper bounds and scores (%) for our two best models, with highest scores bolded.

which one they are referring to.

We hypothesize that these problems contribute to the poor performance of geo-entity linking tools at the city level. We additionally acknowledge how inferring location at the city level is more invasive than inferring at the country or administrative level because it predicts more fine-grained information about the user. Therefore, due to mismatches between geocoordinates and user-defined Location fields as well as the privacy concerns, we suggest that researchers only predict city-level location if it’s necessary for a given application and otherwise use country- or administrative-level predictions.

7 Conclusion and Future Work

In this paper we introduced new methods for geo-entity linking noisy, multilingual social media data with selective prediction. Of our two best performing methods, one does not require any training data (NameGeo+variants) while the other achieves state-of-the-art performance at the country and administrative levels (UserGeo). We also discussed problems with geo-entity linking at the city level for social media data, and suggested against doing this unless necessary for the application.

This work is not without limitations. We do not compare performance with LLM prompting methods, as they are too expensive for large-scale datasets and it’s currently unclear how much prompt testing is necessary for a robust evaluation. However, it would be a useful comparison to know how LLMs perform at zero- or few-shot geo-entity linking; we leave this for future work. We also acknowledge that our methods rely heavily on SBERT models, and thus are reliant on a third party for sustaining them.

We plan to release a version of the UserGeo location embeddings in future, although we will not release the current version due to concerns about the TWITTER-GLOBAL data (e.g. unbalanced across countries, unclear sampling methodology). We also hope to further evaluate our method on other domains with noisy location references, such as

historical data, and to explore extensions of our method so it may be used for the broader task of geoparsing unstructured text.

8 Ethical Considerations

We discuss here ethical issues that may arise from using any geo-entity linking tool on social media data for a downstream application. When using a geo-entity linking tool on social media data, there is always the risk of de-anonymizing users through the inference of sensitive location information. It is generally recommended to use the lowest granularity necessary for the application simply because it is easier to correctly obtain than more fine-grained data (Kruspe et al., 2021), but it is also true that lower geographic granularities protect individuals more due to being part of a larger aggregate (Dupre et al., 2022). We also acknowledge the increasing importance of geomasking techniques, which aim to protect the privacy of individuals while preserving spatial information in geodata. Lorestani et al. (2024) survey the privacy risks of geocoded data and present a taxonomy of current geomasking techniques, and Gao et al. (2019) specifically examine the efficacy of geomasking techniques for protecting the privacy of Twitter users.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments and feedback. This work was supported by a National Science Foundation Graduate Research Fellowship (No. 1938059).

References

- Beatrice Alex, Clare Llewellyn, Claire Grover, Jon Oberlander, and Richard Tobin. 2016. Homing in on twitter users: Evaluating an enhanced geoparser for user profile locations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3936–3944.
- Mariona Coll Ardanuy, Kasra Hosseini, Katherine McDonough, Amrey Krause, Daniel van Strien, and Federico Nanni. 2020. A deep learning approach to geographical candidate selection through toponym matching. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, pages 385–388.
- Mariona Coll Ardanuy and Caroline Sporleder. 2017. Toponym disambiguation in historical documents using semantic and geographic features. In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, pages 175–180.
- Ana Bárbara Cardoso, Bruno Martins, and Jacinto Estima. 2019. Using recurrent neural networks for toponym resolution in text. In *Progress in Artificial Intelligence: 19th EPIA Conference on Artificial Intelligence, EPIA 2019, Vila Real, Portugal, September 3–6, 2019, Proceedings, Part II 19*, pages 769–780. Springer.
- Ana Bárbara Cardoso, Bruno Martins, and Jacinto Estima. 2022. A novel deep learning approach using contextual embeddings for toponym resolution. *ISPRS International Journal of Geo-Information*, 11(1):28.
- Grant DeLozier, Jason Baldrige, and Loretta London. 2015. Gazetteer-independent toponym resolution using geographic word profiles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.
- Catherine D’Ignazio, Rahul Bhargava, Ethan Zuckerman, and Luisa Beck. 2014. CLIFF-CLAVIN: Determining Geographic Focus for News Articles. *NewsKDD: Data Science for News Publishing*.
- Mark Dredze, Michael J Paul, Shane Bergsma, and Hieu Tran. 2013. Carmen: A twitter geolocation system with applications to public health. In *AAAI workshop on expanding the boundaries of health informatics using AI (HIAI)*, volume 23, page 45. Citeseer.
- Sam Dupre, Lindsay Spell, and Paul Jung. 2022. Geospatial data disclosure avoidance and the census. *Select Topics in International Censuses (STIC 1, 1 (2022))*, 1–9.
- Nicholas FitzGerald, Dan Bikel, Jan Botha, Dan Gillick, Tom Kwiatkowski, and Andrew McCallum. 2021. Moleman: Mention-only linking of entities with a mention annotation network. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 278–285.
- Song Gao, Jinneng Rao, Xinyi Liu, Yuhao Kang, Qunying Huang, and Joseph App. 2019. Exploring the effectiveness of geomasking techniques for protecting the geoprivacy of twitter users. *Journal of Spatial Information Science*, (19):105–129.
- Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier. 2018. Which melbourne? augmenting geocoding with maps. Association for Computational Linguistics.
- Claire Grover, Richard Tobin, Kate Byrne, Matthew Woollard, James Reid, Stuart Dunn, and Julian Ball. 2010. Use of the edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1925):3875–3889.
- Bo Han, Paul Cook, and Timothy Baldwin. 2012. Geolocation prediction in social media data by finding

- location indicative words. In *Proceedings of COLING 2012*, pages 1045–1062.
- Xuke Hu, Zhiyong Zhou, Hao Li, Yingjie Hu, Fuqiang Gu, Jens Kersten, Hongchao Fan, and Friederike Klan. 2023. Location reference recognition from texts: A survey and comparison. *ACM Computing Surveys*, 56(5):1–37.
- Binxuan Huang and Kathleen Carley. 2019a. A hierarchical location prediction neural network for Twitter user geolocation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4732–4742, Hong Kong, China. Association for Computational Linguistics.
- Binxuan Huang and Kathleen M Carley. 2019b. A large-scale empirical study of geotagging behavior on twitter. In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 365–373.
- Yuan Huang, Diansheng Guo, Alice Kasakoff, and Jack Grieve. 2016. Understanding us regional linguistic variation with twitter data analysis. *Computers, environment and urban systems*, 59:244–255.
- Mike Izbicki, Vagelis Papalexakis, and Vassilis Tsotras. 2019. Geolocating tweets in any language at any location. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 89–98.
- David Jurgens, Tyler Finethy, James McCorriston, Yi Xu, and Derek Ruths. 2015. Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, pages 188–197.
- Ehsan Kamaloo and Davood Rafiei. 2018. A coherent unsupervised model for toponym resolution. In *Proceedings of the 2018 World Wide Web Conference*, pages 1287–1296.
- Morteza Karimzadeh, Wenyi Huang, Siddhartha Banerjee, Jan Oliver Wallgrün, Frank Hardisty, Scott Pezanowski, Prasenjit Mitra, and Alan M MacEachren. 2013. Geotxt: a web api to leverage place references in text. In *Proceedings of the 7th workshop on geographic information retrieval*, pages 72–73.
- Morteza Karimzadeh, Scott Pezanowski, Alan M MacEachren, and Jan O Wallgrün. 2019. Geotxt: A scalable geoparsing system for unstructured text geolocation. *Transactions in GIS*, 23(1):118–136.
- Anna Kruspe, Matthias Häberle, Eike J Hoffmann, Samyo Rode-Hasinger, Karam Abdulahhad, and Xiao Xiang Zhu. 2021. Changes in twitter geolocations: Insights and suggestions for future usage. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 212–221.
- Sayali Kulkarni, Shailee Jain, Mohammad Javad Hosseini, Jason Baldrige, Eugene Ie, and Li Zhang. 2021. Multi-level gazetteer-free geocoding. In *Proceedings of Second International Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics*, pages 79–88.
- Abhinav Kumar and Jyoti Prakash Singh. 2019. Location reference identification from tweets during emergencies: A deep learning approach. *International journal of disaster risk reduction*, 33:365–375.
- Kathy Lee, Ankit Agrawal, and Alok Choudhary. 2013. Real-time disease surveillance using twitter data: demonstration on flu and cancer. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1474–1477.
- Zekun Li, Jina Kim, Yao-Yi Chiang, and Muhao Chen. 2022. Spabert: A pretrained language model from geographic data for geo-entity representation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2757–2769.
- Zekun Li, Wenxuan Zhou, Yao-Yi Chiang, and Muhao Chen. 2023. Geolm: Empowering language models for geospatially grounded language understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5227–5240.
- Michael D Lieberman and Hanan Samet. 2012. Adaptive context features for toponym resolution in streaming news. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 731–740.
- Zilong Liu, Krzysztof Janowicz, Ling Cai, Rui Zhu, Gengchen Mai, and Meilin Shi. 2022. Geoparsing: Solved or biased? an evaluation of geographic biases in geoparsing. *AGILE: GIScience Series*, 3:9.
- Mahrokh Abdollahi Lorestani, Thilina Ranbaduge, and Thierry Rakotoarivelo. 2024. Privacy risk in geodata: A survey. *arXiv preprint arXiv:2402.03612*.
- Xiangyang Luo, Yaqiong Qiao, Chenliang Li, Jiangtao Ma, and Yimin Liu. 2020. An overview of microblog user geolocation methods. *Information processing & management*, 57(6):102375.
- Ludovic Moncla, Walter Renteria-Agualimpia, Javier Nogueras-Iso, and Mauro Gaio. 2014. Geocoding for texts with fine-grain toponyms: an experiment on a geoparsed hiking descriptions corpus. In *Proceedings of the 22nd acm sigspatial international conference on advances in geographic information systems*, pages 183–192.
- Umashanthi Pavalanathan and Jacob Eisenstein. 2015. Confounds and consequences in geotagged twitter data. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2138–2148.

- Afshin Rahimi, Timothy Baldwin, and Trevor Cohn. 2017. Continuous representation of location for geolocation and lexical dialectology using mixture density networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 167–176.
- Afshin Rahimi, Duy Vu, Trevor Cohn, and Timothy Baldwin. 2015. Exploiting text and network context for geolocation of social media users. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1362–1367.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.
- Hannes Rosenbusch, Anthony M Evans, and Marcel Zeelenberg. 2020. Interregional and intraregional variability of intergroup attitudes predict online hostility. *European Journal of Personality*, 34(5):859–872.
- Breno Dourado Sá, Ticiana Coelho da Silva, and José Antônio Fernandes de Macêdo. 2022. Enhancing geocoding of adjectival toponyms with heuristics. In *Proceedings of the LREC 2022 workshop on Natural Language Processing for Political Sciences*, pages 37–45.
- Praval Sharma, Ashok Samal, Leen-Kiat Soh, and Deepti Joshi. 2023. A spatially-aware data-driven approach to automatically geocoding non-gazetteer place names. *ACM Transactions on Spatial Algorithms and Systems*, 10(1):1–34.
- David A Smith and Gregory Crane. 2001. Disambiguating geographic names in a historical digital library. In *International Conference on Theory and Practice of Digital Libraries*, pages 127–136. Springer.
- Michael Speriosu and Jason Baldridge. 2013. Text-driven toponym resolution using indirect supervision. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1466–1476, Sofia, Bulgaria. Association for Computational Linguistics.
- Jimin Wang and Yingjie Hu. 2019. Enhancing spatial and textual analysis with EUPEG: An extensible and unified platform for evaluating geoparsers. *Transactions in GIS*, 23(6):1393–1419.
- Xiaobin Wang, Chunping Ma, Huafei Zheng, Chu Liu, Pengjun Xie, Linlin Li, and Luo Si. 2019. DM_NLP at semeval-2018 task 12: A pipeline system for toponym resolution. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 917–923.
- Jingyu Zhang, Alexandra DeLucia, and Mark Dredze. 2022. Changes in tweet geolocation over time: A study with carmen 2.0. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 1–14.
- Jingyu Zhang, Alexandra DeLucia, Chenyu Zhang, and Mark Dredze. 2023. Geo-seq2seq: Twitter user geolocation on noisy data through sequence to sequence learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4778–4794.
- Zeyu Zhang and Steven Bethard. 2023. Improving toponym resolution with better candidate generation, transformer-based reranking, and two-stage resolution. *arXiv preprint arXiv:2305.11315*.

A Appendix

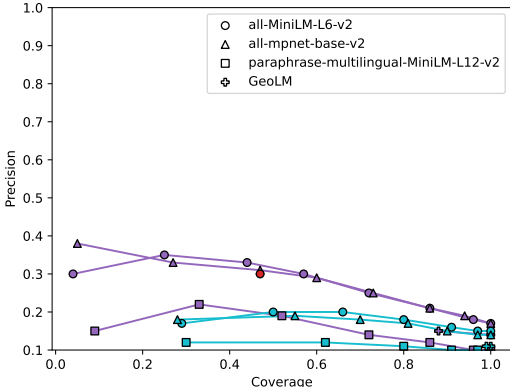


Figure 4: Precision-coverage curve at the city level. The red point is Carmen 2.0, purple points are NameGeo, and cyan are UserGeo.

Country	No. mentions in train data	Carmen 2.0	NameGeo	UserGeo
United States	1300869	.73	.77	.84
Indonesia	713967	.54	.73	.74
Turkey	237472	.57	.91	.91
Malaysia	218257	.49	.70	.70
Brazil	210428	.47	.78	.89
Japan	184647	.27	.48	.92
Philippines	131152	.57	.69	.73
Thailand	70403	.55	.67	.64
Singapore	43996	.51	.54	.53
India	40656	.80	.91	.93
Canada	36990	.54	.83	.78
Saudi Arabia	34559	.40	.48	.73
Argentina	34104	.54	.72	.84
South Africa	27228	.59	.77	.80
Russia	21820	.53	.92	.91
Kuwait	17829	.48	.65	.59
Australia	14363	.53	.85	.86
Chile	14268	.80	.87	.89
Nigeria	12482	.80	.84	.87
Spain	11699	.40	.65	.79
Egypt	10979	.57	.72	.78
UAE	10918	.43	.71	.67
Pakistan	9423	.79	.86	.89

Table 6: Per-country country-level F1-scores for the 23 countries with over 1,000 examples in the test set. Results from NameGeo and UserGeo are using the *all-MiniLM-L6-v2* SBERT model. Both NameGeo and UserGeo outperform Carmen 2.0 for each of the 23 countries. We also note that F1-score can vary significantly by country, demonstrating the importance of evaluating per-country performance when using a geo-entity linker to investigate social media users who are from or identify with particular countries.

News Déjà Vu: Connecting Past and Present with Semantic Search

Brevin Franklin, Emily Silcock, Abhishek Arora, Tom Bryan, and Melissa Dell*
Harvard University, Cambridge, MA, USA. Authors contributed equally.

*Corresponding author: melissadell@fas.harvard.edu.

Abstract

Social scientists and the general public often analyze contemporary events by drawing parallels with the past, a process complicated by the vast, noisy, and unstructured nature of historical texts. For example, hundreds of millions of page scans from historical newspapers have been noisily transcribed. Traditional sparse methods for searching for relevant material in these vast corpora, *e.g.*, with keywords, can be brittle given complex vocabularies and OCR noise. This study introduces News Déjà Vu, a novel semantic search tool that leverages transformer large language models and a bi-encoder approach to identify historical news articles that are most similar to modern news queries. News Déjà Vu first recognizes and masks entities, in order to focus on broader parallels rather than the specific named entities being discussed. Then, a contrastively trained, lightweight bi-encoder retrieves historical articles that are most similar semantically to a modern query, illustrating how phenomena that might seem unique to the present have varied historical precedents. Aimed at social scientists, the user-friendly News Déjà Vu package is designed to be accessible for those who lack extensive familiarity with deep learning. It works with large text datasets, and we show how it can be deployed to a massive scale corpus of historical, open-source news articles. We curate some examples on newsdejavu.github.io. While human expertise remains important for drawing deeper insights, News Déjà Vu provides a powerful tool for exploring parallels in how people have perceived past and present.

"Those who cannot remember the past are condemned to repeat it." – George Santayana's *The Life of Reason*

1 Introduction

Social scientists, and the public more generally, often seek to place the present in perspective by reflecting upon parallels with the past. Finding these

commonalities, however, can be a labor-intensive and challenging process. Vast troves of historical texts have been preserved, but they are often held in unstructured, uncataloged, massive-scale databases. For example, hundreds of millions of pages from historical newspapers have been digitized and are available online through both open-source and proprietary collections. Keyword searches are often used to extract relevant documents from these massive corpora. However, as language is complex and OCR noise is rampant, sparse methods can be extremely brittle.

Transformer large language models offer a powerful tool for retrieving source material from the past that can contextualize the present. This study trains a novel semantic search model, News Déjà Vu, to query which historical news articles are most semantically similar to a modern news article query. Figure 1 shows the model architecture at inference time. Named entities are first detected and masked out, using a named entity recognition model that we tuned for noisy, historical texts. This allows the model to focus on the generalities of the story, rather than specific names of people, locations, or organizations. Then, we use a contrastively trained bi-encoder model to retrieve the modern article's nearest neighbor(s) from a massive-scale database of historical texts.

The News Déjà Vu package allows social scientists to deploy News Déjà Vu on their own queries. It has a CC-BY license and can be used with any appropriately formatted text dataset. It is designed to be user-friendly and intuitive to social scientists, who often lack knowledge of deep learning frameworks. This study provides code snippets showing how it can be used seamlessly with American Stories, a Hugging Face dataset containing over 430 million historical public domain newspaper article texts (Dell et al., 2023). Interested users can query content from a sampling of states in American Stories with modern articles using our

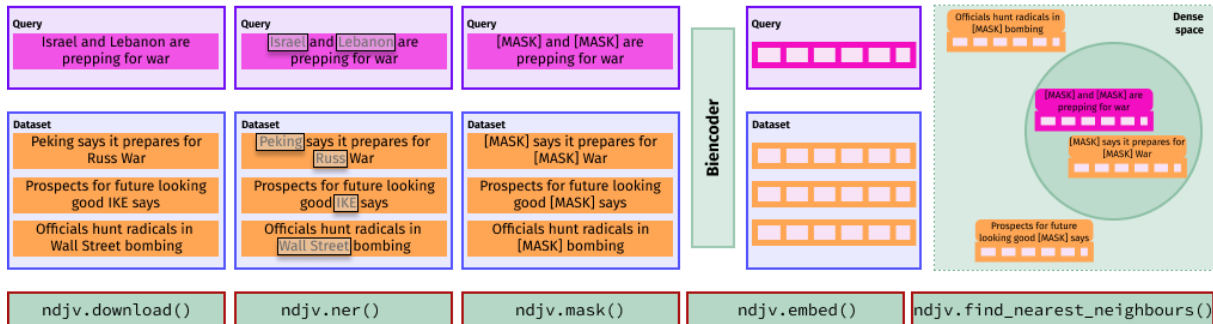


Figure 1: News Déjà Vu architecture at inference time.

HuggingFace demo.¹ We also maintain a website (newsdejavu.github.io), where randomly selected modern news articles - as well as special editions on hand-curated topics of interest - are paired with their retrieved historical neighbors. We built the News Déjà Vu package after many website readers requested that we create a tool that they could use to query their own texts of interest.

News Déjà Vu retrieves articles that use similar semantics. Of course, events or phenomena that are at their core very different may be described in a similar way by the historical and modern news media. This phenomenon is also likely to be of considerable interest to social scientists, but we do caveat that deeper historical knowledge is required to place parallels in appropriate context.

News Déjà Vu currently supports English. In the future, it could be relatively straightforward to create a multilingual model by starting with multilingual Sentence BERT weights and tuning on machine-translated allsides data or other data sources (*e.g.*, [Chen et al. \(2022\)](#)).

The rest of this study is organized as follows. Section 2 discusses the relevant literature, and Section 3 describes the model architecture and training. Section 4 introduces the News Déjà Vu package.

2 Related Literature

There is a large literature on semantic similarity. Most large scale datasets in this space are constructed from web texts. The Massive Text Embedding Benchmark (MTEB) ([Muennighoff et al., 2022](#)), evaluates 8 embedding tasks on 58 datasets covering 112 languages, providing an overview of available datasets.

This study relates most closely to [Silcock et al. \(2022\)](#), which contrastively trains an S-BERT MP-

Net model ([Reimers and Gurevych, 2019](#); [Song et al., 2020](#)) to map historical newswire articles from the same underlying article source to similar embeddings. We initialize News Déjà Vu with their model weights.

More generally, this study follows from the literature on open domain retrieval ([Karpukhin et al., 2020](#); [Thakur et al., 2021](#); [Wu et al., 2019](#)). We also draw inspiration from a large literature showing the importance of contrastive training for semantic similarity applications, which we apply to train News Déjà Vu. The anisotropic shape of the embedding space in pre-trained transformer models like BERT creates challenges for utilizing their latent features ([Ethayarajh, 2019](#)). In these models, less common words are dispersed towards the edges of the hypersphere, the sparsity of low frequency words violates convexity, and the distance between embeddings is correlated with lexical similarity. This leads to misalignment among texts with similar meanings and diminishes the effectiveness of averaging token embeddings to represent longer texts ([Reimers and Gurevych, 2019](#)). By applying contrastive training, anisotropy is mitigated ([Wang and Isola, 2020](#)), enhancing the quality of pooled sentence (or document) representations ([Reimers and Gurevych, 2019](#)).

3 Model Architecture and Training

The News Déjà Vu model architecture at inference time is shown in Figure 1. News Déjà Vu first recognizes and masks spans of text containing named entities (people, locations, organizations, and other miscellaneous proper nouns), as our aim is to draw parallels between articles that describe different entities in different time periods. We then replace all detected entities by the [MASK] token. Query articles are used to retrieve their semantic nearest neighbor(s) in a corpus of interest, using

¹<https://huggingface.co/spaces/dell-research-harvard/newsdejavu>

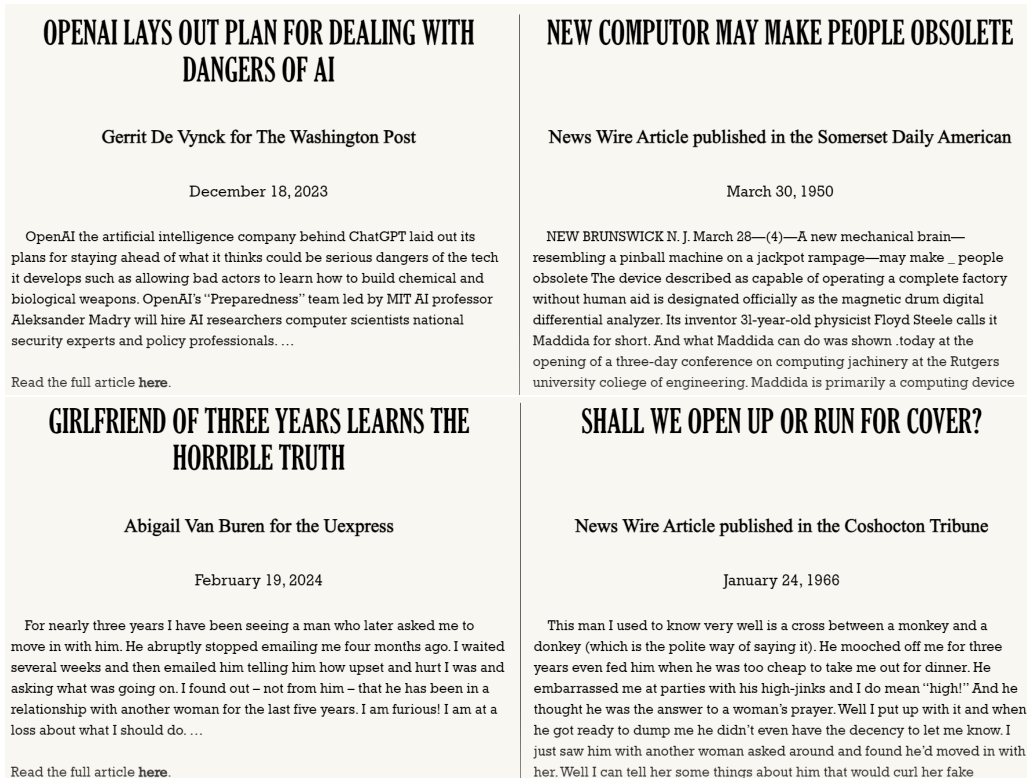


Figure 2: Examples of News Déjà Vu retrieval. The left-hand side shows a modern news article and the right-hand side shows a retrieved historical article.

the News Déjà Vu contrastively trained bi-encoder. We use the IndexFlatIP index from FAISS (Johnson et al., 2019) to perform an exact K-nearest neighbor search. Our embedding vectors are L2 normalized which makes the Inner Product metric used in the index equivalent to Cosine Similarity.

Custom training was necessary to achieve accurate NER performance with robustness to OCR noise. Table 1 describes the training data, which were drawn from randomly selected articles from off-copyright newspaper articles between 1922 and 1977. All data were double-labeled by two highly skilled undergraduate research assistants, and all discrepancies were resolved by hand. The supplemental materials contain the annotator instructions. We use the training set to fine-tune a Roberta-Large model (Liu et al., 2019). We optimised hyperparameters using Hyperband (Li et al., 2018). The best model was trained at a learning rate of $4.7e-05$ with a batch size of 128 for 184 epochs.

Table 2 evaluates NER model performance. This model achieves an F1 of 90.4 in correctly identifying spans of text containing named entities without regards to the class, the relevant task since News Déjà Vu replaces all entities with the [MASK] token. This outperforms a Roberta-Large model

Split	Person	Org	Loc	Misc	Articles
Train	1345	450	1191	1037	227
Val	231	59	192	149	48
Test	261	83	199	181	48

Table 1: The first four columns provide the number of entities of different types in the training, validation, and test sets. The final column provides the total number of labeled articles.

finetuned on CoNLL03 by a large margin.

Model	Precision	Recall	F1
Custom NER	87.9	93.1	90.4
Roberta-Large finetuned on CoNLL03 (Conneau et al., 2019)	80.3	75.5	77.8

Table 2: Evaluation of NER models.

We would like to use News Déjà Vu for unsupervised data exploration, to retrieve historical texts that social scientists and the general public will find thought-provoking. In order to do this, we need an LLM that maps semantically similar articles to similar representations, and we found that off-the-shelf contrastively trained models - such as those described in the MTEB benchmark - did not perform satisfactorily.

Creating paired training data of modern and historical articles that do or do not have parallels would be challenging and costly. Rather, we begin with the model from [Silcock et al. \(2022\)](#), which was contrastively trained on paired historical newswire articles, with the purpose of detecting noisy duplication, rather than semantic similarity. This is a useful starting point since it has already been exposed extensively to the idiosyncrasies of historical news texts, such as OCR errors and obsolete spellings.

We further train on modern data that pairs news articles belonging to the same news story. These are drawn from Allsides, a news aggregator that collates the beginnings of articles on the same story from multiple news sites. Pairs of (the beginnings of) articles from these groupings, which typically consist of two or three texts, form positives. To create negative pairs, we used a larger pool of articles from Allsides, leveraging their pages of articles that are on the same topic, which are broader groupings than those on the same story. We embed this pool using the model from [Silcock et al. \(2022\)](#), which is a finetuned S-BERT MPNET model ([Reimers and Gurevych, 2019](#); [Song et al., 2020](#)). Then for each article that appears in a positive pair, we find the closest article (highest cosine similarity) in the pool that a) is from the same news source and b) does not appear on the same topic or story page.² Training data statistics are given in Table 3.

Training		Validation		Test		Total	
Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.
12868	12913	2757	2766	2757	2766	18,382	18,445

Table 3: Training, validation, and test sizes for the paired data used to train the retriever.

We used Hyperband ([Li et al., 2018](#)) to find optimal hyperparameters, which led us to train for 9 epochs, with a batch size of 32 and a warm-up rate of 0.39. We use S-BERT’s online contrastive loss ([Hadsell et al., 2006](#)) implementation with a margin of 0.5. The model achieves a pairwise F1 of 92.4 on the test set, outperforming models that are not finetuned, such as SBERT MPNET ([Reimers and Gurevych, 2019](#)) and [Silcock et al. \(2022\)](#), as shown in table 4.

We do not have paired modern-historical article texts for evaluation, and it is not clear how one

²In some cases there was no article that met these conditions. In these cases, we took an article from another news source.

Model	Precision	Recall	F1
Custom biencoder	93.7	91.1	92.4
Reimers and Gurevych (2019)	83.8	85.8	84.8
Silcock et al. (2022)	60.7	69.5	64.8

Table 4: Evaluation of biencoder models.

would create such data given the nature of the task. Rather, a skilled annotator gave 350 randomly selected modern articles and their nearest historical News Déjà Vu neighbor a short description of their main topic (typically 2-3 words). In 85.7% of cases, the modern query and its nearest historical neighbor had the same major topic. Even when the major topic is not the same, the pairs often showed other interesting parallels.

Figure 2 provides representative examples of News Déjà Vu retrieval. Other examples can be seen at [newsdejavu.github.io](#). Modern query articles are truncated due to copyright protection. Historical articles are drawn from off-copyright newspapers and reproduced in their entirety. Except for a few special editions about topics requested by our readers, these articles were selected at random, painting a representative picture of News Déjà Vu.

We have a demo where users can use their own texts to query a subset of American Stories ([Dell et al., 2023](#)), a massive scale Hugging Face dataset consisting of over 430 million historical newspaper texts.³ We also make embeddings for the American Stories collection available on Hugging Face.⁴

4 The News Déjà Vu Package

The News Déjà Vu package is available on PyPI for easy install.

```
1 pip install newsdejavu
```

The package consists of the following core functionalities: download, ner, mask, embed, and find_nearest_neighbours. The package focuses on inference, which we expect is how the vast majority of users would like to use News Déjà Vu. For users who wish to fine-tune their own News Déjà Vu model, we recommend using LinkTransformer ([Arora and Dell, 2023](#)) or the Sentence BERT repository ([Reimers and Gurevych, 2019](#)), initializing with our pre-trained weights which are available on Hugging Face.

³<https://huggingface.co/spaces/dell-research-harvard/newsdejavu>

⁴https://huggingface.co/datasets/dell-research-harvard/americanstories_masked_embeddings

The download function downloads the dataset that users would like to work with. We have integrated support for American Stories (Dell et al., 2023). This command allows users to specify which states and year range they would like to download, or they can download the (very large) dataset in full. This step of the pipeline can also be skipped if the user already has a dataset that they would like to use.

```
1 import newsdejavu as ndjv
2 corpus = ndjv.download('american_stories
:1900:Alabama')
```

The ner command runs NER over the corpus.

```
1 ner_outputs = ndjv.ner(corpus, '
historical_newspaper_ner')
```

Next, these detected entities can be masked with a simple mask command and the texts embedded using News Déjà Vu with the embed command. In addition to using the News Déjà Vu model as the default, this command can also support using a local model path or downloading one from Hugging Face, for users who would like to use their own retrieval models in conjunction with the package.

```
1 masked_corpus = ndjv.mask(ner_outputs)
2 corpus_embeddings = ndjv.embed(
masked_corpus, 'same-story')
```

Users can similarly mask and embed their query article. Finally, find_nearest_neighbours retrieves the k closest corpus articles to the query.

```
1 dist_list, nn_list =
find_nearest_neighbours(
query_embeddings, corpus_embeddings,
k=1)
```

Users would typically like to use all these commands in sequence. The mask_and_embed command combines NER, masking, and embedding, and the search_nearest_story command combines NER, masking, embedding, and retrieval.

```
1 corpus_embeddings = ndjv.mask_and_embed(
ner_outputs)
2 nearest_articles = ndjv.
search_nearest_story(query_articles,
'historical_newspaper_ner', 'same-
story', corpus_embed =
corpus_embeddings)
```

We recommend that those who lack extensive familiarity with deep learning frameworks install it on a cloud compute service optimized for deep learning, such as Google Colab, in order to avoid the need to resolve dependencies. Tutorials on how to use News Déjà Vu on Colab will be provided on the News Déjà Vu Github Repository ([github.com/](https://github.com/dell-research-harvard/newsdejavu/)

[dell-research-harvard/newsdejavu/](https://github.com/dell-research-harvard/newsdejavu/)) and on the website.

By making semantic search an accessible tool for social scientists to apply to historical document collections, we hope to make it easier for researchers to find content that contextualizes our understanding of the parallels between past and present.

Ethics Statement

News Déjà Vu is ethically sound. We do emphasize that it retrieves articles that use similar language, which may or may not reflect similarities in the underlying events or phenomena being described. Trained human judgement is required to draw deeper parallels between the past and present, and we hope News Déjà Vu will be a useful tool for directing researchers and the public to content of interest.

Acknowledgements

We are grateful for research assistance from Dennis Du, Jude Ha, Alice Liu, Shiloh Liu, Stephanie Lin, Andrew Lu, Prabhav Kamojjhala, and Ryan Xia.

References

- Abhishek Arora and Melissa Dell. 2023. Linktransformer: A unified package for record linkage with transformer language models. *arXiv preprint arXiv:2309.00789*.
- Xi Chen, Ali Zeynali, Chico Q Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw A Grabowicz, Scott A Hale, David Jurgens, and Mattia Samory. 2022. Semeval-2022 task 8: Multilingual news article similarity.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.
- Melissa Dell, Jacob Carlson, Tom Bryan, Emily Silcock, Abhishek Arora, Zejiang Shen, Luca D’Amico-Wong, Quan Le, Pablo Querubin, and Leander Heldring. 2023. American stories: A large-scale structured text dataset of historical us newspapers. *Advances in Neural Information and Processing Systems, Datasets and Benchmarks*.
- Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*.

- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.
- Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.
- Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Ros-tamizadeh, and Ameeet Talwalkar. 2018. [Hyperband: A novel bandit-based approach to hyperparameter optimization](#). *Preprint*, arXiv:1603.06560.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Man-dar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining ap-proach. *arXiv preprint arXiv:1907.11692*.
- Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.
- Emily Silcock, Luca D’Amico-Wong, Jinglin Yang, and Melissa Dell. 2022. Noise-robust de-duplication at scale. Technical report, National Bureau of Eco-nomic Research.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.
- Nandan Thakur, Nils Reimers, Andreas Rücklé, Ab-hishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evalua-tion of information retrieval models. *arXiv preprint arXiv:2104.08663*.
- Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *International Conference on Machine Learning*, 119:9929–9939.
- Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2019. Scalable zero-shot entity linking with dense entity retrieval. *arXiv preprint arXiv:1911.03814*.

A NER annotator instructions

During the NER annotation process, careful rules were developed to ensure congruence between use of labels. This appendix details those rules

A.1 General rules

- Label the biggest span than constitutes one entity, with the exception of locations. For example “Martin Luther King High School” would be one entity, not “Martin Luther King” or “High School”.
- Label “Vietnamese government”, not just “Vietnamese”.
- “Adam Smith” is one person and the “Catholic Church” is one organization, but “Topeka, Kansas is two locations.
- If the sub-entity is ambiguous without connection to the parent entity, we label it as one entity. For instance, The Department of Electrical Engineering at Tri-State College is all one entity
- Do not include extra punctuation or spaces in the labels, unless they occur within a named entity. If OCR errors produce “First (Baptist Church”, the extra parenthesis should be included since it’s inside a named entity. But if you have “(First Baptist Church”, do not include the parenthesis.
- On a similar note, given something like “Albert Sealy’s cat”, only label “Albert Sealy”. Do not include the apostrophe and the “s”.
- Where relevant, don’t label the “the” (ie. should be “State Department” not “the State Department”).
- For newspapers and other publications/organizations use Google to see if “The” is part of their official name. For instance, “The New York Times” is the actual full name of the newspaper, so label the “The”. However, “The Bell Syndicate” doesn’t actually have “The” as part of their official name, so just label “Bell Syndicate”.
- If there is an organization/location ambiguity, we default to labeling it as a location. If the ‘location’ does an action, it is labeled as an organization. e.g. White House = LOC unless “the White House” does something. e.g. I’m going to the Natural History Museum and the Marriot hotel are both locations.
- EXCEPTION: All administrative units (countries, states) are locations and never organizations, even when doing an action.
- If a location is part of the name of an organization, we label the whole thing as an organization (e.g. “the church of christ in America” or the “Ohio agriculture department”).
- We’re defining locations as 4D: distinctive time periods that are capitalized for places are included in locations e.g. Victorian England, Ancient Greece, Nazi Germany, and Red/Communist China are all labeled as single locations.
- Cardinal direction plus location: if defined in popular speech as a specific location, label the direction also e.g. Western Europe, Central London.
- If the secondary word is plural, it is not a specific location and is not included in the label (e.g. Marriott hotels).
- When established locations are used to reference a location in relation to them, the established locations are labeled separately. E.g. “between Smith and Adams street” are two separate locations.
- This also applies to railroad lines e.g. “the Chicago-New York line” should have Chicago and New York labeled separately.
- Named objects in space (e.g. Sun, Moon, Jupiter) are miscellaneous unless someone/something is going to it or if there is a reference to a place on the object, then they are locations.

A.2 Locations

- If a location title is used as an adjective, it is miscellaneous. e.g. “US senator” should have “US” labelled as MISC, not LOC.

A.3 People

- For person entities, do not label Mr, Mrs, and other prefixes (such as Dr, Rev, etc.) and don’t include suffixes (e.g. “Jr.”, “III”).

- Similarly, do not label descriptors like “Deputy Sheriff” that often come before people’s names.
- Also, positions (without a name) are not named entities e.g.: The Minister of Foreign Affairs, Maharajah, the Queen should not be labelled.
- Include any nicknames when they are put in the middle of names (e.g. something like “Dwayne ‘The Rock’ Johnson” should be all labeled person).
- God is a person, but pronouns like “He” and “Him” are not.
- Animal names are miscellaneous, not person.

A.4 Organizations

- If an organization is used as an adjective, it is miscellaneous.
- Groups of people that go by a name (e.g.: Republicans, Cavaliers) can be orgs, but only if we’re referring to the whole group (e.g. “the republicans” referring to the whole party, not a group of five republicans).
- Liberals and conservatives are not organizations (at least in the US), since the definitions often change and are not really groups.
- Many politician names will be followed by (D/R-State). Label the D/R as miscellaneous, since the group is being used as an adjective to describe the politician. Label the state as location.
- If there is an organization/location ambiguity, we default to labeling it as a location. If the “location” does an action, it is labeled as an organization e.g. ‘the “White House” reported,’ is an organization, but ‘I’m going to the Natural History Museum’ is a location.
- All administrative units (countries, states) are locations and never organizations, even when doing an action.
- Brands of products are organizations if they are the name of the company/producer (e.g. Apple computers) or unless they are now ubiquitous (e.g. ziploc bag, polo shirt). If the brand is not the name of the company, it’s

miscellaneous. E.g.: Toyota is an org, highlander is misc, x-ray is neither, 45-caliper gun is neither.

- Ambiguous organizations are still labeled when they are non-ambiguous in context (e.g. an article talking about Cleveland saying the “rotary club”).
- This also applies to slightly ambiguous government entities like the “army”, “navy”, or “state department” - if it’s clear it’s the US entity in the article, label it as such.

A.5 Miscellaneous

- Adjectives derived from named entities are miscellaneous named entities e.g. nationalities used as adjectives (e.g. U.S., French, London Newspapers).
- EXCEPTION: People used as “adjectives” (e.g. “the Kennedy household”) are to be labeled as PER, not MISC.
- EXCEPTION: When a person’s name has become part of a famous location (e.g. “Eiffel Tower”, “Chandler Building”) that has its own Wikipedia page (or equivalent) the entire location is considered an entity and labeled appropriately.
- When an entity is used as a possessive, that is with an apostrophe or “of” (e.g. “Wisconsin’s cows”, “people of France”), the entity should be labeled with its original label, not MISC.
- Congressional, senatorial, constitutional are all not considered adjectives from named entities.
- Political ideologies are miscellaneous (e.g. communist, socialist, conservative, authoritarian etc.).
- Include prefixes and suffixes to these in misc labels e.g. anti-Japanese, pro-communist.
- Names of groups of people/religions are miscellaneous (only if they’re capitalized or ‘should’ be capitalized, e.g. “communists” should be misc, but “visitors” should not be).
- Titles/positions are not miscellaneous named entities e.g.: The Minister of Foreign Affairs, Maharajah, the Queen are not misc.

- Officially named initiatives/programs are misc e.g. Manhattan Project, U.S. Census.
- Names of capitalized documents or forms are miscellaneous e.g. Individual Census Reports, the Constitution.
- Distinct political acts (e.g. “Agricultural Act of 1970”) are also misc.
- Capitalized/specific names of objects are misc e.g. USS Canopus.
- Names of animals are miscellaneous e.g. Laika (the space dog).
- Events are miscellaneous.
- Events must be famous or distinct (e.g. “Pearl Harbor”, “1969 World Series”).
- Less famous events must refer to something occurring within a specific timeframe (e.g. there will be a “city council meeting” is not a named entity). The event should be unambiguous (e.g. the “city council meeting on 11/5” is not a named entity because it doesn’t specify which city council but the “Boston City Council meeting on 11/5” is).
- Christmas and other major holidays are miscellaneous.
- Brands of products are organizations if they are the name of the company/producer (e.g. Apple computers) or unless they are now ubiquitous (e.g. ziploc bag, polo shirt). If the brand is not the name of the company, it’s miscellaneous. e.g. Toyota is an org, highlander is misc but not org, x-ray is neither, 45-caliper gun is neither.
- Named objects in space (e.g. Sun, Moon, Jupiter) are miscellaneous unless someone/something is going to it or if there is a reference to a place on the object, then they are locations.

B Explanation and Examples for Evaluation of News Déjà Vu

B.1 Explanation

The News Déjà Vu package was used to obtain a total of 5 historical articles from our corpus per modern article of which there were 70. These 70 articles were pulled from the websites of popular news outlets such as the Associated Press, Fox News, USA Today, and the Washington Post. This was done using the News Déjà Vu python package. Each modern article was then placed in a pair with its 5 associated historical articles. A skilled annotator was then tasked with manually classifying the articles in a modern-historical pair as being on or off the same topic.

Being on or about the same topic was defined in the following way: two articles are about the same topic if absent named entities (e.g., proper nouns), the relationship between the most important remaining concepts are essentially the same for both articles. Any two articles that fail to meet these criteria according to the skilled annotator were deemed off or not about the same topic.

Two articles being about the same topic can also be thought about in relation to two articles being on the same story and or on the same events. Topics, stories, and events are distinguished along two dimensions: time and named entities. For two articles to be about the same event, what they each describe must have occurred during the same time period (no more than a day) and have the same or closely related named entities. To be about the same story, two articles must have the same or closely related named entities, but the actions off named-entities or the relationships between named-entities may have occurred over long stretches of time. In both events and stories, named-entities serve the same function as a fictional story's cast of characters. Named-entities and when their actions occur are irrelevant for determining two articles' shared topic. Events, stories, and topics are also nested concepts. Every event is part of a story and every story is part of a topic. An article may also belong to more than one event, story, and/or topic.

An illustrative example might be "Watergate". The "Watergate Break-In" was an event that occurred on Saturday, June 17, 1972 at the Watergate Hotel in Washington, D.C. and was foiled by a security guard named Frank Willis. This is an event because it occurred on a discrete date in time with a clear set of named-entities, like the Watergate Ho-

tel, Washington, D.C., and Frank Willis. If another article was about perpetrator Virgilio Gonzalez's arrest at the Watergate Hotel that night, that article would also be about the "Watergate Break-In" because the arrest occurred at the same date in time. Virgilio Gonzalez is related to the other named entities. The Watergate Hotel in D.C. was the location of his arrest, and it was Frank Willis' tip to the police that led to his arrest. The "Watergate Scandal" would be a story as it unfolded over years and included named entities that were related to one another but did not all participate in the same events. For example, Richard Nixon did not personally break into the Watergate Hotel, but it is believed that he tried to cover up his connections to the break-in's perpetrators, like Virgilio Gonzalez. The cover-up also occurred sometime after the break-in. An article about Richard Nixon's cover-up activities would be about the same story as an article about Virgilio Gonzalez's arrest. "Political Scandals" would be an example of a topic that the "Watergate Scandal" falls under. An article about the "Watergate Scandal" and an article about President Donald Trump pressuring election officials for more votes would both be political scandals. The named entities and dates in time are different, but the most important remaining concepts, like accusations of election interference against presidents, are essentially the same.

B.2 Examples of Historical-Modern Article Pair Evaluation

The following articles make five of the labeled article pairs, each containing the same modern article and one of the five historical articles retrieved using the News Déjà Vu package. Three historical articles were classified as being on the same topic and two were classified as being on different topics. In the same way that you can provide names for events, stories, or topics, like "Political Scandals", the skilled annotator was tasked with doing so for the evaluation of the modern-historical article pairs and those names are included here. In order to abide by copyright restrictions, the full modern article is not reproduced here, just a truncated version. A link to the full article is provided however. Due to OCR errors, the historical article text may appear less coherent than the modern article but still readable nonetheless.

B.2.1 Modern Article

Modern Article URL

<https://www.usatoday.com/story/money/food/2024/03/13/ben-jerrys-free-cone-day-2024/72944410007/>

Modern Article Headline

Ben & Jerry's annual Free Cone Day returns in 2024: Here's when it is and what to know

Modern Article Body

Ben & Jerry's is bringing back its annual Free Cone Day celebration this spring and is asking fans to help them beat a lofty goal.

The company wants this year's Free Cone Day to be the "biggest and best yet with 1 million scoops served," it announced Wednesday.

This year's celebration will take place on Tuesday, April 16, the company said in a news release. Free Cone Day made its return last year after a four-year hiatus.

B.2.2 Historical Article 1

Historical Article ID

10_304477967-ottumwa-daily-courier-Jan-04-1943-p-1.jpg

Historical Article Headline

Additional Cut In Ice Cream Output

Historical Article Body

Washington, D. C.-(P)-The war production board today limited January production on ice cream to 50 per cent of the amount each manufacturer made last October.

- This represented a reduction of one sixth from December, when each manufacturer was permitted to make 60 per cent of his October amount.

The order also applies to frozen custard, milk sherbet, other frozen desserts and ice cream mix.

W.P.B. said the order "was issued at the request of United States department of agriculture . . . » to further relieve the butter shortage."

On Topic with Modern

True

Topic Name / Notes

Ice Cream

B.2.3 Historical Article 2

Historical Article ID

1_106022053-titusville-herald-Apr-11-1939-p-1.jpg

Historical Article Headline

52,299 Jam White House Lawn 'For Annual Easter Ego Rolling

Historical Article Body

OY 2£2RE NAAGCIAER Pear

WASHINGTON, April 10.—Even Alice in Wonderland couldn't have dreamed anything like the Easter rolling which took 52,259 children and adults—by the gate keepers' count—to the private grounds of the White House today.

Through the gates streamed a ten-year-old boy dressed like a white rabbit with flapping pink ears... a live white rabbit scampering on a leash . . . & huge chocolate bunny perched beside a grinning infant in a baby carriage.

Bands played. A magician did tricks. Crowds gathered about a Punch and Judy show. Everywhere, - there were children and eggs—eggs crunched underfoot, smeared on young faces, salling between stately old White House trees and rolling down slopes.

It was like a glorified country picnic. Thousands spread their lunches on the ground. Then it wasn't, for there on the White House porriica wag Lhe President of the United States, waving and wishing, he said, that he could "be down there with you."

And four times during the pleasant sunny day the country's gracious First Lady appeared. Three times she made trips through the grounds, smiling, waving and calling "how-do-you-do" to those along her path. | As she walked she saw children crawl; about on all fours, rolling themselves down slopes, sitting spraddie-legged on the ground to juggle eggs and swingIng from iow limbs of trees.

It was the seventh time, the President pointed out, that he and Mrs. Roosevelt had entertained the children at the White House lawn. The roll has been an annual event at the White House, except for the World war years, since 1878.

"It's a wonderful day," he declared before he disappeared into the house, "and I hope you u enjoy yourselves very, much,?"

On Topic with Modern

False

Topic Name / Notes

one is about ice cream and the other is about Easter

B.2.4 Historical Article 3

Historical Article ID

1_10883667-daily-globe-Mar-01-1949-p-1.jpg

Historical Article Headline

Butter vs. Oleo Battle, Annual Congress Feature, Starts Today

Historical Article Body

Washington — çYfj—The butter vs. oleo battle, an annual feature of congress, got underway today,

Rep. Poage (D-Tex) fired the first shot with an attack on "the butter lobby." He said it is more interested in tryilng "to hill yelJow margarine" than it is in protecting consumers from fraud.

Poage was the first witness at house agriculture committee bhearmgs on 29 bills regarding margarine, Some of them would repeal the taxes on margarine, but prohibit the manufacture und sale of yellow margarine,

Poage sald this latter proposal is an effort en the part of the butter industry "to appear to yield to outraged public sentiment against inexcusable favoritism In favor of one wholesome food

-aguinst anolher without actually lgiving up any af the special pri vilige butler" bas so long enjoy: ed."

POAGE'S BILL

Poage has a bill which would lift, federal taxes on margarine' and permit manufacturers 10 col: or it yellow. But it would require public eating places to notify cus: tomers \$f they serve margarine.

The issue as whether there should be a federal tax on the

l butter substitute. The government ;Now taxes all of it, with on ex. tra toll if at is ecalored

President Truman for repeal of the margarine ti ;@s, and so did the Democri platform on which he ran list fall, The Republican platform did ; not mention the subject.

Tf the controversy follows its new almost traditional form — butter partisans will suggest that margarine could be colored anything from purple or green to a bright, cherry red=anything but yellow,

The margarine folks will say butter has no exclusive cham to yellow, and in fact the yellow color has to be added ta some bulter.

Southern Democrats from rural districts and northern Democrats sand even Republicans — from big city districts will ally for the moment in support of marparine,

There are two types of Dills before ihe commiltee,

The most numerous would abolish all federal taxes on margnrine. Others would but out st least the added nx on colored margare.

At the other end of the line are bis by Reps. August 11. An ;aresen (R-Minn), und Granger '(D-Uiah) to prohibit the manufacture and sule of yellow marigaring, STARTED 69 YEARS AGO

The whole thing started 63 years ago, when the first federal anti-margarine Jaw was pussed. That law and others which followed have been under attack almost every year since.

The oleo forees made their greatest progress last yeat, when a bill fo repeal the margarine taxes passed the house handily. It never got to a vote in the sen l ate.

A federal tax of 10 cents a pound is now paid on all colored margarine sold at retail. The retail tax on the uncolored product is one-quarter cent a pound.

Margarine manufacturers pay a federal tax of \$600 a year; wholesalers pay a tax of \$480 4 year Sf they handle colored margarine, and \$200 a yenr for uncolored, and retailers are taxed \$48 n year fo handle colored mare garine and \$6 10 handle uncol

On Topic with Modern

True

Topic Name / Notes

food

B.2.5 Historical Article 4

Historical Article ID

26_90193365-morning-herald-Apr-04-1946-p-1.jpg

Historical Article Headline

No More KP In The Air Forces Of Army

Historical Article Body

Washington, April 3 (P) There will be no more KP (Kitchen Police) duty in the Army Air Forces under a new program announced today.

Soldiers will still peel spuds and wash dishes, But those who do will be permanently assigned to the task and will be called "mess attendants." The announcement adds that they "will be afforded an opportunity to make an Army career of food service,"

The old system of assigning all men on the roster to KP in turn is being abolished.

The AAF announcement said that "many (local exigencies and personnel problems" prevent setting a definite date for the establishment of Utopia.

On Topic with Modern

True

Topic Name / Notes

Food

B.2.6 Historical Article 5

Historical Article ID

5_233435376-circleville-herald-Dec-30-1976-p-1.jpg

Historical Article Headline

International Falls No Sunny Spa

Historical Article Body

INTERNATIONAL FALLS, Minn. (AP) — At the close of a year, a time for reflection, hardly a place is better suited than this for that worthy exercise.

There is something to be learned from this place, something other than what everybody already knows from the nightly weather report: that it is the coldest place in the 48 contiguous states.

When winter's fangs bite into this little spot on the Canadian border — in the first half of this month the thermometer managed to hang above zero for only four brief hours — living becomes an adventure and humility a daily lesson. Nature's elemental severity invites long thoughts about man's standing in the Great Scheme.

"I think we worry more about the simple necessities of survival than most people do," said Frank Bohman, a philosophic aviator who has lived here all his 52 years.

"Having enough food in the house, enough fuel, a backup heating system, these are real concerns. I would imagine that in gentler climates people take survival for granted."

For the record, when the earth tilts toward winter, winds borne on the jet stream sweep from the North Pole down the interior flank of the Canadian Rockies and pivot eastward right at this point, so that the average yearly temperature here is 37.5 degrees and the annual snowfall 50 inches. Readings in the minus 30s and 40s are commonplace during the winter.

The cold grips so fiercely, in fact, that it all but refuses to let go. The ground freezes five feet down, untillable until June.

The town is on the granite shore of Rainy Lake, one of creation's masterpieces, a 340-square-mile work of art done in a freeform of coves and bays and flecked with 1,600 tiny granite islands timbered with pine.

Thus in the summertime the area is awash with tourists, regulars who return to their summer places on the islands, weekenders seeking walleyed pike and clean air, visitors with cameras, water skis and time to make the two-hour drive up from Duluth.

When the summer crop of frolickers is harvested, however, only a bold band of the hearty remain to face the long dark winter.

That yearly experience has given them a palpable sense of neighborliness, a closeness such as a shared secret brings. When the ice breaks up each May they have earned a communal handshake that says nice going everybody, we did it again, we didn't quit.

Those brave souls number 9,109 in. International Falls and the nearby communities of South International Falls and Ranier. About the same number

live across the Rainy River Bridge at Fort Frances, Canada.

The spirit of hands-across-the-sea, or in this case, the river, comes naturally; nature's legacy knows no international boundary. Indeed, one longtime Chamber of Commerce president in International Falls, Gordy McBride, was a Canadian citizen.

On Topic with Modern

False

Topic Name / Notes

one about ice cream and the other about weather

Knowledge Distillation in Automated Annotation: Supervised Text Classification with LLM-Generated Training Labels

Nicholas Pangakis and Samuel Wolken
University of Pennsylvania
{njpang@sas., sam.wolken@asc.}upenn.edu

Abstract

Computational social science (CSS) practitioners often rely on human-labeled data to fine-tune supervised text classifiers. We assess the potential for researchers to augment or replace human-generated training data with surrogate training labels from generative large language models (LLMs). We introduce a recommended workflow and test this LLM application by replicating 14 classification tasks and measuring performance. We employ a novel corpus of English-language text classification data sets from recent CSS articles in high-impact journals. Because these data sets are stored in password-protected archives, our analyses are less prone to issues of contamination. For each task, we compare supervised classifiers fine-tuned using GPT-4 labels against classifiers fine-tuned with human annotations and against labels from GPT-4 and Mistral-7B with few-shot in-context learning. Our findings indicate that supervised classification models fine-tuned on LLM-generated labels perform comparably to models fine-tuned with labels from human annotators. Fine-tuning models using LLM-generated labels can be a fast, efficient and cost-effective method of building supervised text classifiers.

1 Introduction

Supervised text classification often relies on human-labeled text data for training and validation. Computational social science (CSS) researchers frequently use these types of supervised models to classify large quantities of text, ranging from news articles on the internet to government documents (Grimmer et al., 2022; Lazer et al., 2020). Collecting training and validation labels generated by humans for these tasks, however, is expensive, slow, and prone to a variety of errors (Grimmer and Stewart, 2013; Neuendorf, 2016).

To address these limitations, prior research suggests utilizing few-shot capabilities of generative

large language models (LLMs) to annotate text data instead of human annotators (Gilardi et al., 2023). Generative LLMs are faster and cheaper than human annotators and do not suffer from common human challenges such as limited attention span or fatigue. While this approach has its limitations and generative LLMs do not excel at all text annotation tasks (Pangakis et al., 2023), prior research illustrates that there are numerous circumstances where generative LLMs can produce high quality text-annotation labels.¹

Although past work suggests LLM few-shot annotation is highly effective, it may be cost prohibitive in many settings. Research with text data often involves classifying millions of documents or text samples. For example, a recent CSS article studies a data set of 6.2 million tweets labeled on four dimensions (Hopkins et al., 2024), a task that would have cost nearly \$9,000 if using GPT-4 alone.² Using a knowledge distillation approach (Dasgupta et al., 2023; Gou et al., 2021; Hinton et al., 2015), it may be possible to approximate the performance of a larger “teacher” model (e.g., GPT-4 (OpenAI, 2023), estimated to have over 1.7T parameters (Schreiner, 2023)) with much smaller and cheaper task-specific “student” models (e.g., BERT Base (Devlin et al., 2019), approximately 110 million parameters).

In this paper, we evaluate using generative LLMs to create surrogate labels for fine-tuning downstream supervised classification models. Our approach involves first using a generative LLM to label a subset of text samples and then fine-tuning supervised text classifiers with the LLM-generated labels. Using our outlined approach, we replicate 14 classification tasks from recently published CSS articles. We compare several supervised classifiers (i.e., BERT (Devlin et al., 2019), RoBERTa (Liu

¹See Appendix A.1 for a longer discussion of automated annotation research in CSS.

²Appendix A.2 elaborates on costs with LLM annotation.

et al., 2019), DistilBERT (Sanh et al., 2019), XLNet (Yang et al., 2020), and Mistral-7B (Jiang et al., 2023)) fine-tuned on varying quantities of either human-labeled samples or GPT-4-labeled samples. We benchmark the supervised classifiers’ performance against GPT-4 and Mistral-7B few-shot labels. In a series of ablation experiments, we also explore whether GPT-4 outputs change over time and how well the student models handle noise in the GPT-generated text labels.

A small number of studies have utilized similar approaches in related domains. Chen et al. (2023b) use ChatGPT annotations to train various Graph Neural Networks for a fraction of the cost of human annotations. Golde et al. (2023) also harness ChatGPT to create surrogate text data that aligns with a specific valence (i.e., positive and negative) and then subsequently fine-tune a supervised classifier using the synthetic text. Most analogous to our approach here, Wang et al. (2021) train RoBERTa (Liu et al., 2019) and PEGASUS (Zhang et al., 2020) models on labels generated by GPT-3. Despite strong performance across their analyses, Wang et al. (2021), as well as the previously mentioned studies, exclusively evaluate closed-source models (i.e., GPT-3 and ChatGPT) on popular, publicly available NLP benchmark tasks (e.g., AG-News, DBpedia, etc), which are plausibly included in the training data for the generative LLM. As a result, these analyses cannot offer a clear indication of performance because their results plausibly suffer from contamination (Balepur et al., 2024; Li and Flanigan, 2023; Magar and Schwartz, 2022; Srivastava et al., 2024). Put otherwise, strong performance may reflect memorization, which casts doubt on the generalizability of the findings.

To compare supervised classifiers fine-tuned using LLM-generated labels against those fine-tuned with labels from human annotators, researchers must assess performance on tasks less likely to be affected by contamination. To this end, all 14 of the classification tasks we replicate are conducted on labeled data sets stored in password-protected archives. Each of the classification tasks in our corpus are real CSS applications and contain human-labeled ground-truth annotations.³

Our main contributions are as follows:

1. Across 14 classifications tasks, supervised models fine-tuned with GPT-generated labels

³Table A2 and Table A3 include a full list of the data sets and classification tasks.

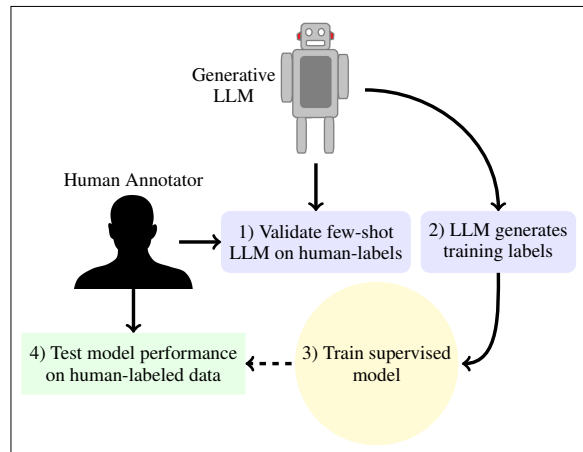


Figure 1: Supervised text classification with LLM-generated training labels.

perform comparably to models fine-tuned with human-labeled data. The median F1 performance gap between models fine-tuned using GPT-labels and models fine-tuned on human-labeled data is only 0.039. While supervised classifiers fine-tuned with LLM-generated labels perform slightly worse than classifiers fine-tuned with human labels, LLM-generated labels can be a fast, efficient and cost-effective method to fine-tune supervised text classifiers.

2. Supervised models fine-tuned on GPT-4 generated labels perform remarkably close to GPT few-shot models, with a median F1 difference of only 0.006 across the classification tasks.
3. GPT-4 few-shot models and supervised classifiers fine-tuned on GPT-4 generated labels perform significantly better than all other models on *recall*, but noticeably worse on *precision*.

2 Methodology

Figure 1 shows our four-step workflow. First, we validate LLM few-shot performance against a small subset (n=250) of human-labeled text samples for each task. We provide GPT-4⁴ with detailed instructions to label the text samples into conceptual categories outlined in the original study.⁵ Because LLM few-shot annotation performance varies across tasks and data sets, validation is always necessary (Pangakis et al., 2023). As such, we validate

⁴We select GPT-4 as our main generative model due to its high performance on popular leaderboard websites. In Appendix E.1, we also explore few-shot performance of an open-source model (i.e., Mistral-7B).

⁵We include all prompt details in the supplementary material. We also include our code to query the GPT-4 API.

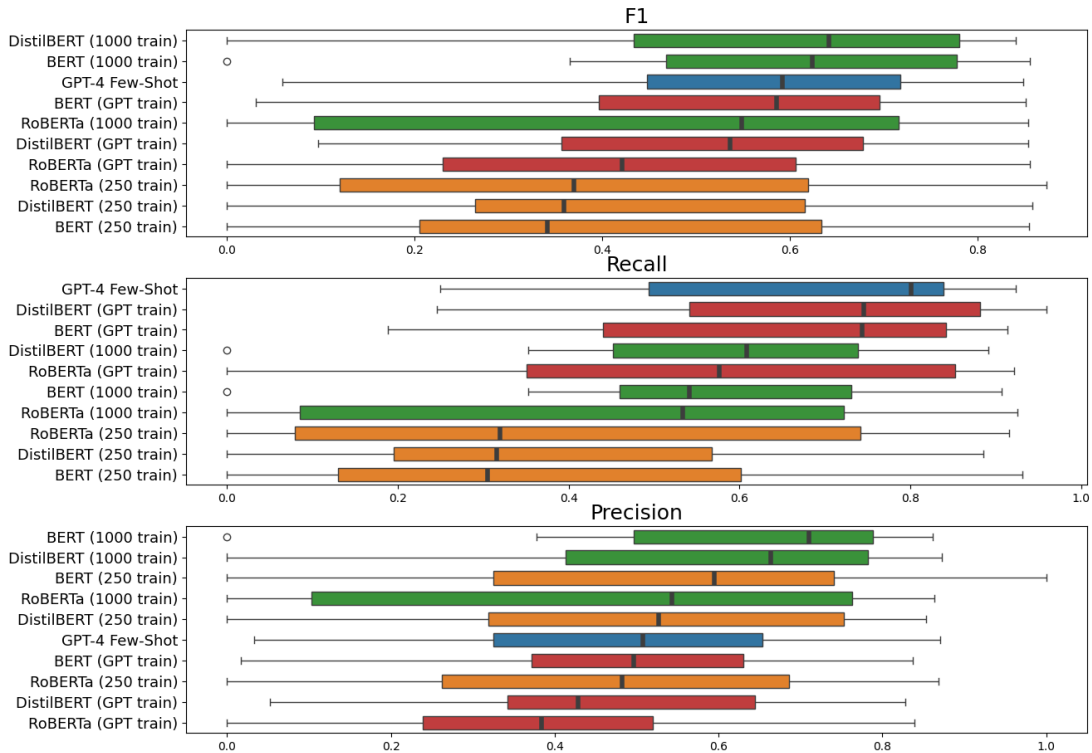


Figure 2: Box plots of performance on test data across 14 tasks. Thick vertical line denotes median. Color represents model type, with green corresponding to models fine-tuned on 1,000 human labels, orange to 250 human labels, red to 1,000 GPT labels, and blue to a few-shot model.

each generative LLM on a subsample and then adjust the prompt to optimize performance on this initial sample. This process is discussed in greater detail in Appendix C.1. Using the validated prompt, the second step in our workflow involves labeling an additional 1,000 text samples per task using the same generative LLM, which will later be used as data to fine-tune the supervised classifier.

In the third and fourth steps, we fine-tune a variety of supervised text classifiers and assess performance against a held-out set of 1000 human-labeled samples. Our supervised models include a variety of BERT-family models (i.e., BERT, RoBERTa, and DistilBERT).⁶ In Appendix E.1, we conduct ablation experiments with XLNet and Mistral-7B. Appendix C.2 describes on our hyperparameter tuning process and additional evaluation details, including how multi-class tasks were split into separate binary tasks. Ultimately, we compare performance between text classifiers fine-tuned on 1000 LLM-generated samples, 250 human-labeled samples, and 1000 human-labeled samples.

In addition to analyzing performance across dif-

⁶We select these models because of their low cost, speed, and their frequent application in CSS (Büyüköz et al., 2020; Terechshenko et al., 2020).

ferent model architectures and training sample sizes, we also implement a variety of ablation experiments to assess how robust the analyses are to several sources of variance. First, we examine how robust these models are to noisy GPT-generated labels. Specifically, in Appendix E, we implement a novel technique designed to measure noise in GPT-generated labels and then compare supervised models fine-tuned on GPT-generated labels *with noise* against models fine-tuned on GPT-generated labels *without noise*. In a second set of ablation experiments, we replicate the GPT-4 few-shot labels at different points in time. To account for the potential of changing model weights in GPT-4, we re-analyzed each task six months after our initial analyses and compared results across time. Extended discussion and the results for these ablation experiments are shown in Appendix E.

3 Results

Classification results for the BERT-family models and GPT-4 few-shot are shown in Table 1.⁷ In Figure 2, each box plot displays the range of

⁷We conduct few-shot classification by using the classification instructions from the original study as a prompt for the LLM.

Model	Training data	Accuracy	F1	Precision	Recall
GPT-4	Few shot	0.88	0.59	0.51	0.80
BERT	Human annotation: 250	0.89	0.34	0.59	0.30
	Human annotation: 1000	0.92	0.62	0.71	0.54
	GPT-4 annotation: 1000	0.87	0.59	0.50	0.74
DistilBERT	Human annotation: 250	0.89	0.36	0.53	0.32
	Human annotation: 1000	0.89	0.64	0.66	0.61
	GPT-4 annotation: 1000	0.85	0.54	0.43	0.75
RoBERTa	Human annotation: 250	0.88	0.37	0.48	0.32
	Human annotation: 1000	0.90	0.55	0.54	0.53
	GPT-4 annotation: 1000	0.84	0.42	0.38	0.58

Table 1: Comparison of classification performance on held-out validation data. Median performance across 14 tasks shown.

evaluation metrics across all 14 tasks for a given model/training data combination. The thick vertical line denotes the median performance metric across all analyzed tasks. Across all 14 classification tasks, DistilBERT and BERT fine-tuned on 1000 human-samples are the highest performing models, with a median F1 score of 0.641 and 0.624, respectively.⁸ Not far behind, however, is the GPT-4 few-shot model (0.592 median F1) and BERT fine-tuned on 1000 GPT-labeled samples (0.586 median F1). From this we draw two conclusions: First, models fine-tuned on few-shot surrogate labels from a generative LLM perform comparably to models fine-tuned on human labels. Despite a small performance gap, training supervised models on LLM-labeled data can be a quick, effective, and budget-friendly approach for constructing supervised text classifiers.

Second, models trained on surrogate labels from GPT-4 demonstrate very similar validation performance as labels from GPT-4 with few-shot in-context learning. As each additional GPT-4 query incurs more expense, researchers can save resources by avoiding classifying an entire data set using a generative LLM and instead use them to create training labels for a supervised model.

A secondary finding is that GPT few-shot models and supervised models trained on GPT-generated labels produce remarkably high performance on recall.⁹ GPT-4 few-shot (0.8 median recall) as well as DistilBERT and BERT fine-tuned on GPT-labels

(both with 0.746 median recall) achieve significantly better median recall than any model fine-tuned with human labels. The opposite is true for precision: BERT fine-tuned on human-labels achieved the highest precision of the models tested, which was 0.214 higher than median precision for BERT models fine-tuned on GPT-4 labels. Therefore, using surrogate training labels may be better suited for tasks where recall is prioritized over precision.

4 Discussion

Surrogate labels from generative LLMs offer a viable, low-resource strategy for fine-tuning task-specific supervised classifiers, but a few points of caution are worth emphasizing. As the variation in our few-shot results indicates, there are cases where GPT-4 performs poorly on classification tasks. While advancements in LLM technology and additional prompt engineering could mitigate these concerns, it is essential that researchers validate generative LLM performance against ground-truth human-labeled data. Downstream supervised classifiers will not mitigate bias or poor performance in LLM few-shot labels. Thus, while generative LLMs can improve the classification workflow, their application must remain human-centered.

⁸We use F1 as our primary evaluation criteria due to class imbalance. Full results are shown in Table A4.

⁹Appendix D displays PR curves for each of the BERT-family supervised models.

5 Limitations

Here, we identify three main limitations of our analysis. First, as discussed in Section 4 and shown in full detail in Table A4, there are various circumstances where supervised models fine-tuned on LLM-generated labels fail to produce satisfactory results. This may be due to inaccurate annotations from GPT-4, poor performance from the supervised classifier, or both. While it is possible that additional prompt engineering or hyperparameter tuning could improve performance, it is essential to stress that each of these optimization strategies rely on human labels for comparison. As a result, we argue that it is essential to center human judgement as ground truth when optimizing models and adjudicating between models.

A second, related limitation refers to understanding the errors in the model outputs. Specifically, it is possible that errors from a GPT-trained model produces correlated but unobservable errors. Building a supervised classifier on top of GPT-4 labels would magnify, rather than offset, any such biases. This, too, underscores the importance of human validation and error analysis. It is, of course, also essential to minimize bias by human annotators. For instance, recruiting human annotators from varying demographic backgrounds when conducting an annotation project may diminish the potential for correlated errors across annotators.

Finally, treating human labels as ground truth is an additional limitation. Although most data sets in our analysis employed multiple human coders, it is of course possible that these annotators made correlated errors. As a result, some disagreements between human ground truth labels and surrogate GPT-4 labels may stem from human error. Such errors could bias performance metrics downward for any of the models assessed. Because our primary interest is making comparisons across models, however, we are mainly interested in their relative performance. Because each model would suffer from the same errors in the human labeled data, we do not see this as a significant concern for this analysis.

For the analysis in this paper, our reliance on text classification tasks and data from peer-reviewed research in high-impact journals helps to mitigate concerns about data annotation quality. The annotation procedures in each of these tasks received IRB approval and was assessed by independent reviewers to be of quality enough for publication in

a high-impact journal. Still, it is important to acknowledge that applied researchers should invest in high-quality human labels, even if only to validate generative LLM annotation performance.

6 Ethics Statement

Our research complies with the ACL Ethics Policy. Specifically, our research positively contributes to society and human well-being by providing tools that can aid computational social scientists studying the social world. Using the methods we introduce and test will help scientists better understand a wide range of complicated social problems. Because the techniques proposed and assessed in this article require dramatically less resource expenditure than alternatives, our results can help address inequities in resources across researchers.

Due to the inherent risks of deploying biased models, we stress the necessity of human validation throughout our paper. Given the ease and efficiency gains of using generative LLMs to train supervised classifiers, we believe it is essential to build rigorous testing and evaluation standards that are human-centered. This is why we took great efforts to center our analyses on data sets less prone to contamination risks.

Moreover, our research and data analysis does not cause any harm while also respecting privacy and confidentiality concerns. As we discuss in our data collection procedures in Appendix B, we conformed to each data repository’s usage and replication policies. Each of the original studies received IRB approval and our analyses conformed to the same safety protocols. All collected data was anonymized by the original authors. Appendix C.3 provides additional details on human annotation protocols, which were all conducted by the original studies and received IRB approval.

References

- Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. 2024. [Artifacts or abduction: How do llms answer multiple-choice questions without the question?](#)
- Berfu Büyüköz, Ali Hürriyetoglu, and Arzucan Özgür. 2020. Analyzing elmo and distilbert on socio-political news classification. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pages 9–18.
- Dallas Card, Serina Chang, Chris Becker, Julia Mendelsohn, Rob Voigt, Leah Boustan, Ran Abramitzky, and

- Dan Jurafsky. 2022. [Computational analysis of 140 years of us political speeches reveals more positive but increasingly polarized framing of immigration](#). *Proceedings of the National Academy of Sciences of the United States of America*, 31.
- Lingjiao Chen, Matei Zaharia, and James Zou. 2023a. [How is chatgpt’s behavior changing over time?](#)
- Zhikai Chen, Haitao Mao, Hongzhi Wen, Haoyu Han, Wei Jin, Haiyang Zhang, and Hui Liu and Jiliang Tang. 2023b. [Label-free node classification on graphs with large language models \(llms\)](#).
- Cheng-Han Chiang and Hung-yi Lee. 2023. [Can large language models be an alternative to human evaluations?](#) *arXiv preprint arXiv:2305.01937*.
- Michael Chmielewski and Sarah C. Kucker. 2020. [An mturk crisis? shifts in data quality and the impact on study results](#). *Social Psychological and Personality Science*, 11(4):464–473.
- Sayantana Dasgupta, Trevor Cohn, and Timothy Baldwin. 2023. [Cost-effective distillation of large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7346–7354, Toronto, Canada. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, page 4171–4186.
- Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq Joty, and Boyang Li. 2022. [Is gpt-3 a good data annotator?](#)
- Benjamin D. Douglas, Patrick J. Ewell, and Markus Braue. 2023. [Data quality in online human-subjects research: Comparisons between mturk, prolific, cloudresearch, qualtrics, and sona](#). *PLoS One*, 18.
- Naoki Egami, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2022. [How to make causal inferences using texts](#). *Science Advances*, 8(42):eabg2652.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. [Chatgpt outperforms crowd-workers for text-annotation tasks](#).
- Jonas Golde, Patrick Haller, Felix Hamborg, Julian Risch, and Alan Akbik. 2023. [Fabricator: An open source toolkit for generating labeled training data with teacher llms](#).
- Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. 2021. [Knowledge distillation: A survey](#). *International Journal of Computer Vision*, page 1789–1819.
- Justin Grimmer, Margaret E. Roberts, and Brandon Stewart. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press.
- Justin Grimmer and Brandon M. Stewart. 2013. [Text as data: The promise and pitfalls of automatic content analysis methods for political texts](#). *Political Analysis*, 21(3):267–297.
- Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2023. [Annollm: Making large language models to be better crowdsourced annotators](#).
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. [Distilling the knowledge in a neural network](#).
- Daniel J. Hopkins, Yphtach Lelkes, and Samuel Wolken. 2024. [The rise of and demand for identity-oriented media coverage](#). *American Journal of Political Science*.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth e Lacroix, and William El Sayed. 2023. [Mistral 7b](#).
- Ross Deans Kristensen-McLachlan, Mical Canavan, M rton Kardos, Mia Jacobsen, and Lene Aar e. 2023. [Chatbots are not reliable text annotators](#).
- David MJ Lazer, Alex Pentland, Duncan J Watts, Sinan Aral, Susan Athey, Noshir Contractor, Deen Freelon, Sandra Gonzalez-Bailon, Gary King, and Helen Margetts. 2020. [Computational social science: Obstacles and opportunities](#). *Science*, 369(6507):1060–1062.
- Changmao Li and Jeffrey Flanigan. 2023. [Task contamination: Language models may not be few-shot anymore](#).
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#).
- Inbal Magar and Roy Schwartz. 2022. [Data contamination: From memorization to exploitation](#).
- Wes McKinney. 2011. pandas: a foundational python library for data analysis and statistics. *Python for high performance and scientific computing*, 14(9):1–9.
- Jonathan Mellon, Jack Bailey, Ralph Scott, James Breckwoldt, and Marta Miori. 2022. [Does gpt-3 know what the most important issue is? using large language models to code open-text social survey responses at scale](#). Working paper.

- Stefan Müller. 2022. The temporal focus of campaign communication. *The Journal of Politics*, 84(1):585–590.
- Kimberly A. Neuendorf. 2016. *The Content Analysis Guidebook*. Sage Publications.
- OpenAI. 2023. [Gpt-4 technical report](#).
- Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. 2023. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. *Proceedings of the 40th International Conference on Machine Learning*, pages 26837–26867.
- Nicholas Pangakis, Samuel Wolken, and Neil Fasching. 2023. [Automated annotation with generative ai requires validation](#).
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, and Trevor Killeen et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*.
- Hao Peng, Daniel M. Romero, and Eموke-Agnes Horvat. 2022. Dynamics of cross-platform attention to retracted papers. *Proceedings of the National Academy of Sciences*, 119(25):585–590.
- Michael Reiss. 2023. [Testing the reliability of chatgpt for text annotation and classification: A cautionary remark](#). Working paper.
- Christopher Michael Rytting, Taylor Sorensen, Lisa Argyle, Ethan Busby, Nancy Fulda, Joshua Gubler, and David Wingate. 2023. [Towards coding social science datasets with language models](#).
- Punyajoy Saha, Narla, Komal Kalyan, and Animesh Mukherjee. 2023. [On the rise of fear speech in online social media](#). *Proceedings of the National Academy of Sciences of the United States of America*.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. [Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter](#).
- Maximilian Schreiner. 2023. [Gpt-4 architecture, datasets, costs and more leaked](#). Blog post.
- Saurabh Srivastava, Annarose M B, Anto P V au2, Shashank Menon, Ajay Sukumar, Adwaith Samod T, Alan Philipose, Stevin Prince, and Sooraj Thomas. 2024. [Functional benchmarks for robust evaluation of reasoning performance, and the reasoning gap](#).
- Zhanna Terechshenko, Fridolin Linder, Vishakh Padmakumar, Michael Liu, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. 2020. A comparison of methods in political science text classification: Transfer learning language models for politics. *Working Paper*.
- Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. From humans to machines: can chatgpt-like llms effectively replace human annotators in nlp tasks. *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media*.
- Petter Törnberg. 2023. [Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning](#).
- Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. [Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks](#).
- Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. [Want to reduce labeling cost? gpt-3 can help](#).
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison and Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. *In Proceedings of EMNLP*.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. [Xlnet: Generalized autoregressive pretraining for language understanding](#).
- Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *International Conference on Machine Learning*, page 11328–11339.
- Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. [Can chatgpt reproduce human-generated labels? a study of social computing tasks](#).
- Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. [Can large language models transform computational social science?](#) Working paper.

A Appendix: Prior automated annotation research in computational social science

A.1 Overview of automated annotation research

A growing body of research studying automated annotation claims that few-shot classifications from generative LLMs can match humans on annotation tasks (Chiang and Lee, 2023; Ding et al., 2022; Gilardi et al., 2023; He et al., 2023; Mellon et al.,

GPT-4: Entire Corpus (n=6.2m)	GPT-4: n=1000	Crowdworker: n=1000	Trained Assistant: n=1000
\$8,990	\$15	\$124	\$187

Table A1: Comparing annotation costs applied to Hopkins et al. (2024).

2022; Pan et al., 2023; Rytting et al., 2023; Thapa et al., 2023; Törnberg, 2023; Zhu et al., 2023; Ziems et al., 2023). For example, Gilardi et al. (2023) find that LLMs outperform typical crowd-sourced human annotators: “[t]he evidence is consistent across different types of texts and time periods. It strongly suggests that ChatGPT may already be a superior approach compared to crowd annotations on platforms such as MTurk.” Analyzing a range of social science applications, Rytting et al. (2023) similarly write, “GPT-3 can match the performance of human coders [and in] some cases, it even outperforms humans in increasing intercoder agreement scores.” Törnberg (2023) argues that automated annotations by LLMs in his analyses are even as accurate as annotations by human experts. While there are clearly circumstances where automated annotation fails to accurately reflect human judgment (Kristensen-McLachlan et al., 2023; Reiss, 2023), researchers can safely use automated annotation procedures as long as they validate against human labels not prone to contamination (Pangakis et al., 2023).

A.2 Costs associated with implementing automated annotation

While prior research demonstrates that automated annotation can align with human reasoning in many scenarios, directly using the strategies introduced in prior studies to label an entire text corpus would be cost prohibitive when applied to a typical CSS data set, which often contain millions of observations. Consider the cost for using GPT-4 to label a data set of 6.2 million tweets, which is what Hopkins et al. (2024) analyze. At the time of writing, GPT-4 costs \$0.01 per 1k input tokens and \$0.03 per 1k output tokens, with 1000 tokens corresponding to roughly 750 words.¹⁰ The prompt instructions to replicate Hopkins et al. (2024) contained approximately 500 words and the average tweet length was around 25 words. Because the full corpus contained 6.2 million tweets and the code to query the OpenAI API was implemented in batches of 10 tweets, a full automated annotation to process

the corpus in Hopkins et al. (2024) would require 620,000 batches fed into GPT-4. Each batch (i.e., 750 words per input) corresponds to roughly 1,000 input tokens, per OpenAI’s suggested benchmark. Since the outputs were standardized, the outputs for these analyses tended to be around 150 tokens.

Thus, when broken down into tokens, the total number of processed input tokens for this analysis would be $1,000 \times 620,000$ and the total processed output tokens would be $150 \times 620,000$. When factoring the cost per token for input and output tokens, the total cost comes to $\$8,990 = (1,000 \times 620,000 \times 0.00001) + (150 \times 620,000 \times 0.00003)$. While this is a loose estimate, it illustrates the challenges posed by the marginal per-sample cost of automated LLM annotation for large-N CSS research. Using our approach, labeling 1,000 text samples and training a supervised classifier would cost under \$15.

Implementing our proposed workflow also reduces annotation labor costs. For example, hiring crowd-source workers to label a subset of text samples to serve as training observations would still cost significantly more than using automated annotation. Hopkins et al. (2024), for example, hire MTurk workers and paid them \$0.06 to \$0.07 per task depending on the total number of annotations (\$15.00 per hour for six tasks per minute), which extrapolates to 360 tasks per hour. Under the standard assumption of three MTurk workers per task and taking a majority vote, the entire annotation time to label 1,000 tweets would have taken slightly under three hours and cost \$124. However, due to serious data quality concerns about crowdworkers (Chmielewski and Kucker, 2020; Douglas et al., 2023; Veselovsky et al., 2023), a better cost comparison is against trained research assistants instead. Assuming 45 seconds per task and a \$15 hourly rate, manually annotating 1,000 text samples would take 12.5 hours and cost approximately \$187.

Table A1 shows a comparison of these costs. Not only is automated annotation remarkably faster than human annotators, our procedures introduced here can cost researchers less than 10% the cost of typical alternatives. These efficiency gains are

¹⁰See <https://openai.com/pricing>

conservative in the sense that they disregard the time to find, hire, and train the annotator.¹¹

B Appendix: Data sets

In this section, we elaborate on the data sets used in our analysis. Our corpus includes 14 classification tasks across five data sets representing recent applications in computational social science. To avoid the potential for contamination, we rely exclusively on data sets stored in password-protected data archives (e.g., Dataverse). We draw from research published in outlets across a spectrum of disciplines ranging from interdisciplinary publications (e.g., *Proceedings of the National Academy of Sciences*) to high-impact field journals in social science (e.g., *American Journal of Political Science*). To find these articles, we searched journals for articles related to computational social science that implemented some type of manual annotation procedure. The human-labeled data from the original study is treated as the ground truth. We discuss the human annotation procedures in the original studies at greater length in Appendix C.3.

It is important to note that while the raw data (e.g., tweets and Facebook posts) may be included in the LLM pretraining data, the accompanying labels from the human annotators are certainly not included in the pretraining data. This is because the labels accompanying each text sample (e.g., whether a tweet referenced a specific racial identity frame) are not public-facing. If the text without the associated label is not included in the pretraining data, there is no cause for concern that the annotation task would suffer from contamination.

Table A2 and Table A3 contain the full details for every task and data set. Overall, our data encompass diverse degrees of class imbalance: Across tasks, the mean positive class frequency is 16.2%, the minimum is 0.04%, and the maximum is 61%. The sources of labels are representative of common approaches to annotation: 42.9% of tasks were annotated by crowdsourced workers, 28.6% by experts, and 28.6% by research assistants.

Our replications involve fine-tuning supervised classifiers using manually annotated data from the replication data sets. For every replication clas-

sification task, we conformed to each data repository’s replication policies. Each of the original studies received IRB approval and our analyses conformed to the same safety protocols, including full anonymization and agreeing to not publicly post the raw data without permission. As such, our replication of each data set is compatible with its intended usage.

Although all of the data sets were anonymized before our replications, we manually reviewed each data set to confirm privacy protections. One of the data sets (Saha et al., 2023) contains hate speech, but this is because it is a central part of the research question from the original study. As a result, we include examples of hate speech in that particular replication. From manual review, no other data set contained offensive material.

C Appendix: Additional methodological details

C.1 Prompt tuning

As discussed in Section 2, for every task we adjusted each GPT-4 prompt with a human-in-the-loop update procedure to optimize for accurate annotations. This human-in-the-loop process involved three steps. First, we used the generative LLM to annotate a small subset of the text samples per task ($n=250$).¹² Second, we manually reviewed instances where humans and the generative LLM disagreed on the text’s label. Because our accuracy at this stage hovered around 0.8, this usually entailed manually reviewing roughly 50 text labels. Third, we adjusted the prompt instructions to clarify instances where automated annotation failed to correctly align with human judgment.

The prompt tuning process should be minimal (e.g., one or two iterations), because any further efforts could lead to overfitting the prompt to a small subset of the data (Egami et al., 2022). If the prompt is overly tailored to a small subset of the data, then the instructions may not generalize to unseen data. Moreover, if the researcher makes major changes to the prompt, there may be a mismatch between the human annotator’s codebook and the generative LLM’s instructions. Like the previous concern, the differences in the instructions could lead to poor performance on a held-out set. As a result, if there are substantial changes made to the LLM’s prompt, then the researcher

¹¹It is worth stressing here that validation against human-created labels is still essential. Therefore, researchers may want to prioritize their budgets for hiring domain experts to code a small subset of data to serve as validation and test data, as we demonstrate in Figure 1. Our cost efficiency calculations are based on training data, not validation and test sets.

¹²This subset of text samples was not included in the held-out test set.



Figure A3: Change in LLM annotation performance on training data after one round of prompt optimization

should also change the human codebook as well and re-annotate new text samples. As such, these procedures should not be resource or time intensive. Instead, prompt tuning is intended to be a part of a validation process of few-shot in-context learning.

Some researchers argue that small changes to the LLM prompt instructions can dramatically alter automated annotation performance (Reiss, 2023), whereas others claim that alterations have a marginal effect (Rytting et al., 2023). To test how variations in the prompt instructions affect performance, we evaluated automated annotation performance before and after the prompt tuning process.

Figure A3 shows the distributions of change in performance metrics after updating the LLM prompt and re-annotating the same text samples. This analysis demonstrates whether and how prompt optimization affects LLM annotation, holding constant the data and conceptual categories. In most cases, prompt optimization led to minor improvement in accuracy and F1—although recall decreased in more cases than improved after updating the prompts. The small magnitude of change in classification performance suggests that generative LLMs are fairly robust to slight word changes in the prompt, which aligns with prior work that conducts similar experiments (Rytting et al., 2023). While the magnitude of improvement was generally small, researchers experiencing subpar LLM annotation performance can use human-in-the-loop prompt optimization to ensure that their instructions are not the cause of poor performance.

Qualitatively, the most common mistakes we ob-

served by the generative LLM during the prompt optimization stage were false positives stemming from the text sample containing language broadly associated with the conceptual category of interest. For example, one task focused on identifying immigration content in American political speeches (Card et al., 2022). Initially, the generative model consistently categorized a text sample as containing an immigration reference if the speech mentioned a foreign country or foreign national, irrespective of whether the mention was connected to immigration in any way. For the prompt-update process for this task, changes in this case meant clarifying that any reference to a foreign country or foreign national did not warrant a positive class instance unless it was explicitly referenced in relation to American immigration or immigration policy. While this process was manual, we also believe that future work could conduct these procedures algorithmically—plausibly using generative AI as well.

C.2 Hyperparameter tuning, evaluation, and compute details

Our experiments involved varying the training data used to fine-tune numerous supervised classifiers (i.e., 250 human samples, 1000 human samples, and 1000 GPT-labeled samples). To select each supervised classifier, we implemented a grid search over 18 possible hyperparameter combinations. In particular, we optimized learning rate (1e-5, 2e-5, and 5e-5), batch size (8 and 16), and epochs (2, 4, and 6). We conducted our search on a subsample of 250 text samples per task and retained the best hyperparameters (in terms of highest F1) across each task. We subsequently used the best-performing

combination of hyperparameters for all applications of a specific model (see best-performing hyperparameter configurations in Table A5). Despite not adopting a more exhaustive approach to hyperparameter tuning, we observe strong performance across our classification tasks, with a few exceptions. Table A6 displays additional model hyperparameters that remained constant across tasks, as well as basic information about each model’s architecture.

Overall, for each task we had a total of 2,500 labeled text samples labeled by both human annotators and the LLM: (1) a training set of 1,000 text samples; (2) two separate validation sets (both with $n=250$); and (3) a test set ($n=1000$). Each of these sets of data were labeled by humans and the generative LLM. The training set ($n=1000$) was used to fine-tune the supervised classifiers. The first validation set ($n=250$) was used to optimize the generative LLM prompt and validate its few-shot performance. The second, separate validation set ($n=250$) was used to conduct our grid search. The test set ($n=1000$) was used to assess the final performance of the few-shot model and the supervised models.

For all 14 tasks, evaluation was conducted on a test set of 1000 held-out text samples that had previously been labeled by human annotators. To harmonize the diverse range of annotation tasks into a common framework for evaluation, we treat every task dimension as a separate binary annotation task. Thus, if an article included a classification task with three potential labels, we split the annotation process into three discrete binary classification tasks. As is standard in binary classification evaluation, we report accuracy, F1, precision, and recall for every task and model.¹³ Table A4 displays the full classification results across all tasks and models.

All of our supervised training analyses were implemented in Python 3.10.12 with HuggingFace’s Transformers (Wolf et al., 2020) and PyTorch libraries (Paszke et al., 2019). We conducted all data preprocessing in Python Pandas (McKinney, 2011). Our computing infrastructure was Google Colab, where we used 215 T4 GPU compute units (roughly 421.4 GPU hours). As with our model selection, we chose this computing environment due to its low cost and ease of application. Any computational social scientist could conduct the same analyses. In

¹³Because our tasks are all binary, there is no need to report any multi-label classification metrics, like Macro-F1.

the supplementary material, we include all code to run our supervised training procedures.

C.3 Additional details on human annotation procedures

We introduce a novel corpus of labeled text data for annotations. To create this data set, we compile labeled data from recent studies, as detailed in A2. As a result, we did not work with annotators to generate any original data. We adopted materials from these original studies instead. While we do not report the instructions given to each study’s human annotators, we do provide the prompt instructions that were used to query GPT-4 in the supplementary material. These instructions were taken directly from the original study’s human annotator instructions. All additional details on the annotation procedures (e.g., how they were recruited, payment, consent, and demographic characteristics) can be found in the original studies’ supplementary material.

While we do not describe each study’s procedures in detail, we manually selected our annotation studies due to their high-quality human labeling practices. All of the replicated studies were approved by an IRB. These studies all deployed either expert coders or numerous non-expert coders of varying backgrounds. Because all of the human annotation text is part of the peer-review process in high-impact journals and due to the strict annotation guidelines and principles these studies adhered to, we conclude that the human annotations are of high-quality.

D Appendix: Extended results

Figure A4 shows precision-recall (PR) curves for each of the BERT-family models trained on either human labels or GPT labels, pooling all classification tasks. The decrease in performance for GPT-generated labels compared with human labels is small based on area under the curve (AUC). Thus, supervised classifiers trained with GPT-generated labels perform comparably to classifiers trained with human-generated labels on these tasks. Across models and tasks, precision appears to drop below 1.0 around 0.7 recall.

E Appendix: Ablation experiments

We conducted a variety of ablation experiments to account for sources of variance. The next three

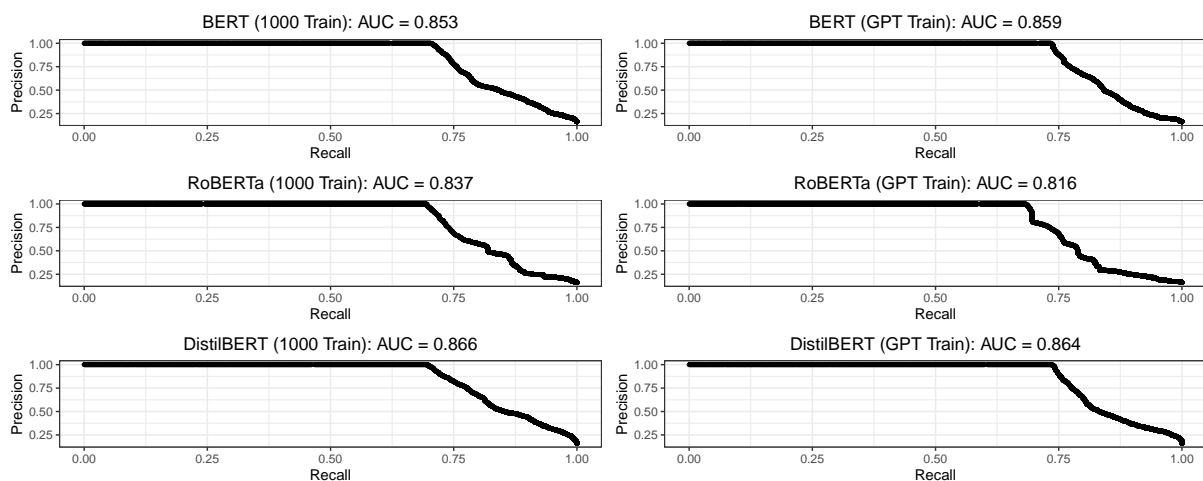


Figure A4: Precision-recall curves across each BERT-family model

sections detail these experiments and their main findings.

E.1 Comparing classifiers with different model size and architecture

First, to account for variation in model architecture and model size, we compare performance across two additional language models for supervised classification (i.e., XLNet and Mistral-7B). These models are beyond the BERT-family models included in the main analyses (i.e., BERT, DistilBERT, and RoBERTa). In addition to a Mistral-7B supervised sequence classification model, we also generate few-shot labels using Mistral-7B using the same procedures we employed in the GPT-4 few-shot model.

The primary difference between the BERT-family models and XLNet is the training objective. The BERT-based models are pretrained using a Masked Language Modeling (MLM) objective, whereas XLNet is an autoregressive model that uses Permutation Language Modeling (PLM), which involves learning context across input tokens in any permutation order. In addition to being significantly larger than the BERT models, Mistral-7B utilizes a distinct type of attention in the pretraining process (i.e., grouped-query attention (GQA) and sliding window attention (SWA)). We include the Mistral-7B few-shot model as a smaller, open-source alternative to GPT-4. Mistral-7B was selected because the model weights are available for download and it displays higher performance than Llama-13B (Jiang et al., 2023).

Figure A5 shows the classification performance

from these additional models and compares them to the results from BERT and GPT-4 few-shot in the main analyses. The test set for these analyses is the same as the main analysis shown in the paper. Our results from examining these additional models do not change the substantive conclusions in the paper: Models trained on surrogate training labels perform comparably to models trained with human labeled data. XLNet even performs slightly better than the fully human labels. The gap between Mistral-7b fine-tuned using human labels and GPT-labels, however, is notably larger than the other models, with a median difference of 0.12. Overall, BERT and GPT-4 still appear to be the strongest performing models.

There is also a fairly sizeable gap between the open-source (Mistral-7B) and closed-source (GPT-4) few-shot models. Although it may be expected from a significantly smaller and free-to-use model, F1 scores for Mistral-7B are 0.16 worse, on average, than GPT-4. Mistral-7B also took significantly longer to run than GPT-4. These findings further reinforce the necessity of human validation.

E.2 Comparing classifiers with and without noise

Our second set of ablation experiments involve comparing supervised models trained on GPT-generated labels *with noise* against GPT-generated labels *without noise*. To measure noise in the GPT-labels, we utilize the predicted token sampling process of generative LLMs to gauge an LLM’s “confidence” in the annotation of each text sample. By introducing randomness in the LLM sampling

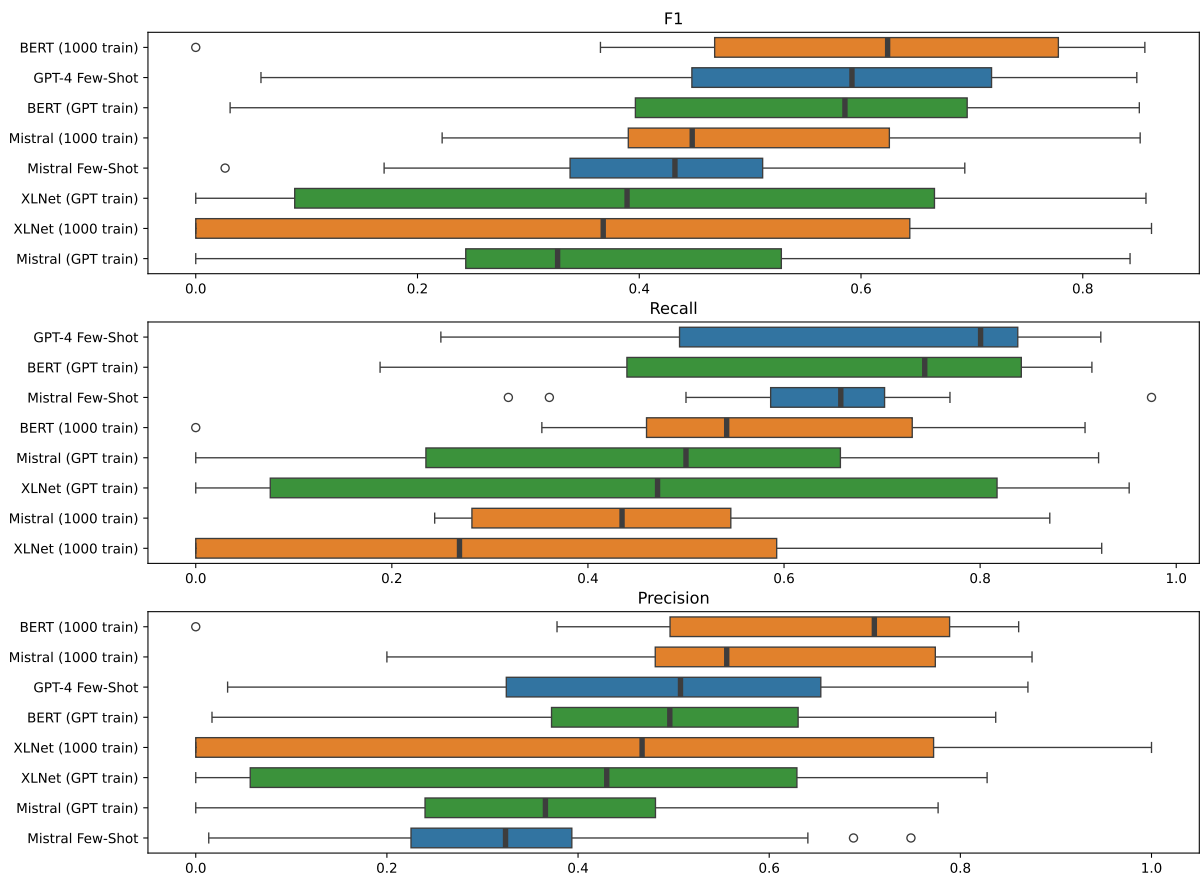


Figure A5: Box plots of ablation performance on test data across 14 tasks. Thick vertical line denotes median.

process through the temperature setting and by repeatedly classifying the same text sample multiple times, we identify text samples that cannot be clearly classified into one of the annotation categories specified by the prompt instructions.¹⁴

Classifications that vary across iterations may be “edge cases” and have a lower probability of correct classification.¹⁵ This approach rests on the core assumption that the full distribution of token probabilities captures latent information about the annotation’s classification. If, for example, the top tokens are similar in probability, then choosing one of these tokens may misrepresent the model’s annotation decision. Instead, measuring the variability across iterations allows us to find these “edge cases.” We call this measure of uncertainty in the annotation label a “consistency score.” We define an indicator function $C(i)$ that is equal to 1 when label i for a given task is equal to the LLM’s modal classification, m , for task :

$$C(i) = \begin{cases} 1 & \text{if } i = m \\ 0 & \text{otherwise} \end{cases}$$

Given a vector of classifications, \mathbf{a} , with length l for a given classification task, *consistency* is measured as the proportion of classifications that match the modal label:

$$Consistency = \frac{1}{l} \sum_{j=1}^l C(a_j)$$

For these ablation experiments, we classify every text sample three times at a temperature of 0.7 and measure each text sample’s consistency score. Because there are only three iterations, each text sample can only have two values for consistency score: 0.67 and 1.0. Across all analyzed tasks, classifications with a consistency of 1.0 show significantly higher accuracy (19.4% increase), true positive rate (16.4% increase), and true negative rate (21.4% increase) compared to classifications with a consistency less than 1.0. Roughly 85% of classifications had a consistency of 1.0.

¹⁴Generative LLMs output a series of probabilities that correspond to each token in its vocabulary. To select a specific token from this probability distribution, generative LLMs sample a randomly selected token, weighted by its probability. The temperature hyperparameter governs this sampling process. A higher temperature setting flattens the probability distribution and causes the sampling draw to become more uniform across tokens. A lower temperature, however, isolates the sampling to select only the most likely tokens.

¹⁵Accessing token log probabilities directly, once available, will be an effective way to a similar analysis.

Table A7 shows supervised model performance for BERT models fine-tuned on 1,250 training observations labeled by GPT-4 (i.e., labels with noise) compared to BERT models fine-tuned on training observations with a consistency score of 1.0 (i.e., labels without noise), which reduced our training set to slightly more than 1000 samples per task. Put otherwise, the second series of models involved dropping about 250 text samples per task so that the training set only retained annotations where GPT-4 consistently labeled the same category across all iterations.

Our findings indicate that there are minimal differences between models trained on labels with noise and labels without noise. Models trained without noise display, on average, 0.004 lower F1 score than models trained with noise. These results suggest that the supervised models explored here are fairly robust to noise in the labels.

E.3 Comparing GPT-4 few-shot performance over time

Our final set of ablation experiments involved replicating the GPT-4 few-shot model at different points in time. An unsettling scenario involves the potential drift in capabilities as generative LLMs undergo opaque changes and updates. Some research, such as Chen et al. (2023a), claim that GPT-4 performance is declining over time. To account for the potential of changing model weights in GPT-4, we re-analyzed each task six months after our initial analyses and compared results across time.

Figure A6 shows evaluation comparisons of few-shot tasks in both April 2023 and November 2023. Our results do not suggest significant changes in GPT-4 performance over time. If anything, Figure A6 reveals a small *increase* in performance since my initial experiments. Across the 14 tasks, accuracy improved by 0.007 and F1 increased by 0.022 when the same annotation procedures were carried out in November 2023.

F Appendix: Miscellaneous additional information

Additional sources:

- Robot image (used in Figure 1): https://commons.wikimedia.org/wiki/File:Grey_cartoon_robot.png
- Human silhouette image (used in Figure 1): https://commons.wikimedia.org/wiki/File:SVG_Human_Silhouette.svg

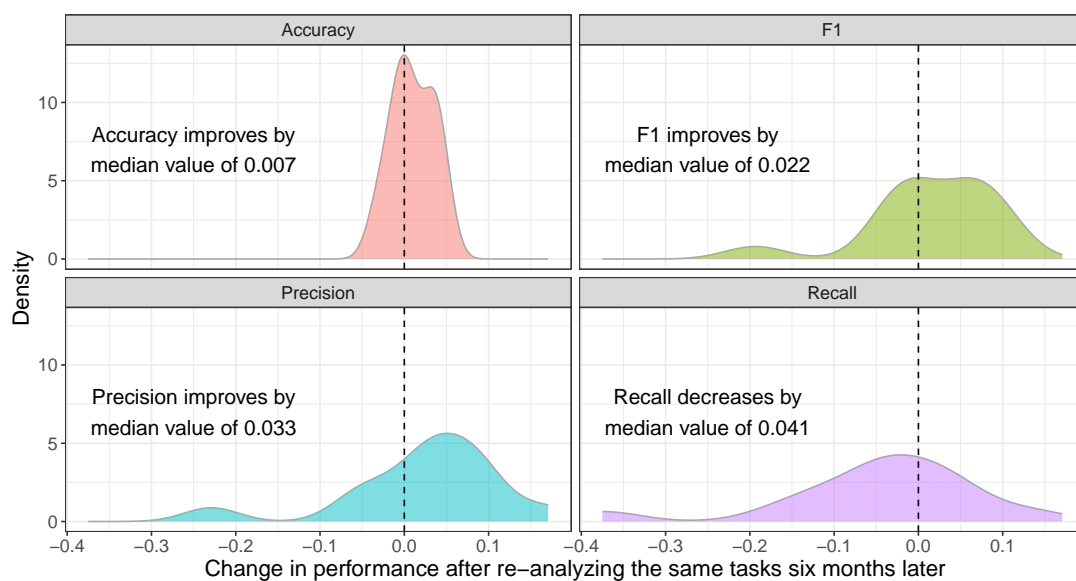


Figure A6: Examining GPT-4 performance over time

Author(s)	Title	Journal	Year
Card et al.	Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration	PNAS	2022
Hopkins, Lelkes, and Wolken	The Rise of and Demand for Identity-Oriented Media Coverage	American Journal of Political Science	2024
Müller	The Temporal Focus of Campaign Communication	Journal of Politics	2021
Peng, Romero, and Horvat	Dynamics of cross-platform attention to retracted papers	PNAS	2022
Saha et al.	On the rise of fear speech in online social media	PNAS	2022

Table A2: Replication data sources.

Study	# of tasks	Annotation source	Classification tasks
Card et al. (2022)	4	Research assistants	Classify US congressional speeches to identify whether the speech discussed immigration or immigration policy, along with an accompanying tone: pro-immigration, anti-immigration, or neutral.
Hopkins, Lelkes, and Wolken (2024)	4	Crowd	Classify headlines, Tweets, and Facebook share blurbs to identify references to social groups defined by a) race/ethnicity; b) gender/sexuality; c) politics; d) religion.
Müller (2021)	3	Expert	Classify sentences from political party manifestos for temporal direction: past, present, or future.
Peng, Romero, and Horvat (2022)	1	Expert	Classify whether Tweets express criticism of findings from academic papers.
Saha et al. (2020)	2	Crowd	Classify social media posts into fear speech, hate speech, both, or neither.

Table A3: Descriptions of replication classification tasks.

Data set	Task	Model	Training data															
			Few shot				Human: 250				Human: 1000				GPT: 1000			
			Ac.	F1	Pr.	Re.	Ac.	F1	Pr.	Re.	Ac.	F1	Pr.	Re.	Ac.	F1	Pr.	Re.
Card et al.	Cat: Neg	GPT-4	0.85	0.65	0.54	0.83	0.88	0.58	0.74	0.48	0.87	0.56	0.65	0.49	0.81	0.56	0.47	0.72
		BERT					0.85	0.51	0.59	0.45	0.84	0.48	0.55	0.42	0.78	0.57	0.43	0.82
		RoBERTa					0.86	0.56	0.61	0.51	0.86	0.58	0.61	0.55	0.81	0.58	0.47	0.74
		DistilBERT																
	Cat: Imm	GPT-4	0.81	0.81	0.74	0.90	0.85	0.84	0.79	0.89	0.86	0.86	0.81	0.91	0.84	0.83	0.76	0.91
		BERT					0.86	0.85	0.80	0.92	0.85	0.84	0.77	0.92	0.82	0.82	0.74	0.92
		RoBERTa					0.85	0.84	0.80	0.88	0.84	0.84	0.79	0.89	0.82	0.82	0.73	0.92
		DistilBERT																
	Cat: Neut.	GPT-4	0.83	0.26	0.27	0.25	0.80	0.35	0.29	0.44	0.85	0.36	0.38	0.35	0.87	0.38	0.44	0.34
		BERT					0.88	0.30	0.46	0.23	0.88	0.00	0.00	0.00	0.84	0.33	0.33	0.34
		RoBERTa					0.85	0.28	0.32	0.25	0.85	0.36	0.37	0.35	0.86	0.38	0.40	0.36
		DistilBERT																
Cat: Pro	GPT-4	0.88	0.50	0.55	0.46	0.86	0.33	0.44	0.27	0.84	0.44	0.42	0.46	0.87	0.45	0.51	0.40	
	BERT					0.87	0.37	0.51	0.30	0.84	0.37	0.41	0.34	0.85	0.41	0.43	0.39	
	RoBERTa					0.87	0.29	0.55	0.19	0.83	0.38	0.38	0.37	0.84	0.35	0.40	0.31	
	DistilBERT																	
Hopkins et al.	Political	GPT-4	0.88	0.43	0.30	0.79	0.95	0.32	0.60	0.22	0.96	0.62	0.71	0.54	0.82	0.34	0.21	0.82
		BERT					0.84	0.37	0.23	0.85	0.96	0.62	0.73	0.54	0.84	0.37	0.23	0.85
		RoBERTa					0.94	0.29	0.50	0.20	0.96	0.63	0.72	0.56	0.83	0.34	0.22	0.80
		DistilBERT																
	Gender	GPT-4	0.95	0.74	0.68	0.82	0.91	0.20	0.46	0.13	0.96	0.80	0.86	0.74	0.94	0.72	0.62	0.85
		BERT					0.91	0.08	0.44	0.04	0.95	0.73	0.78	0.68	0.92	0.67	0.54	0.87
		RoBERTa					0.94	0.52	0.83	0.38	0.97	0.81	0.87	0.75	0.93	0.71	0.59	0.88
		DistilBERT																
	Race	GPT-4	0.96	0.57	0.41	0.92	0.97	0.00	0.00	0.00	0.98	0.56	0.71	0.46	0.98	0.64	0.54	0.77
		BERT					0.97	0.00	0.00	0.00	0.97	0.00	0.00	0.00	0.97	0.59	0.45	0.85
		RoBERTa					0.97	0.00	0.00	0.00	0.99	0.71	0.77	0.65	0.97	0.54	0.46	0.65
		DistilBERT																
Religion	GPT-4	0.98	0.61	0.47	0.88	0.98	0.21	1.00	0.12	0.99	0.73	0.75	0.71	0.98	0.61	0.48	0.82	
	BERT					0.98	0.00	0.00	0.00	0.98	0.00	0.00	0.00	0.98	0.00	0.00	0.00	
	RoBERTa					0.98	0.00	0.00	0.00	0.99	0.69	0.67	0.71	0.97	0.53	0.37	0.94	
	DistilBERT																	
Müller	Future	GPT-4	0.82	0.85	0.87	0.83	0.83	0.85	0.88	0.84	0.82	0.85	0.85	0.85	0.81	0.85	0.84	0.87
		BERT					0.84	0.87	0.87	0.88	0.82	0.85	0.86	0.85	0.82	0.86	0.84	0.87
		RoBERTa					0.83	0.86	0.85	0.86	0.81	0.84	0.87	0.82	0.82	0.85	0.83	0.88
		DistilBERT																
	Past	GPT-4	0.91	0.74	0.66	0.84	0.94	0.83	0.74	0.93	0.95	0.83	0.80	0.85	0.93	0.79	0.71	0.89
		BERT					0.94	0.80	0.81	0.79	0.95	0.85	0.79	0.92	0.85	0.00	0.00	0.00
		RoBERTa					0.94	0.79	0.77	0.80	0.94	0.80	0.79	0.82	0.93	0.79	0.68	0.96
		DistilBERT																
	Present	GPT-4	0.82	0.62	0.64	0.60	0.83	0.65	0.66	0.64	0.83	0.65	0.64	0.66	0.81	0.61	0.63	0.58
		BERT					0.84	0.66	0.71	0.61	0.84	0.68	0.68	0.67	0.83	0.61	0.68	0.56
		RoBERTa					0.83	0.64	0.69	0.59	0.83	0.65	0.66	0.64	0.82	0.59	0.66	0.54
		DistilBERT																
Peng et al.	Critical	GPT-4	0.85	0.54	0.48	0.63	0.87	0.43	0.59	0.34	0.91	0.63	0.76	0.54	0.79	0.43	0.35	0.56
		BERT					0.88	0.44	0.61	0.34	0.87	0.62	0.54	0.73	0.78	0.43	0.34	0.59
		RoBERTa					0.83	0.43	0.42	0.44	0.86	0.54	0.50	0.58	0.77	0.41	0.33	0.56
		DistilBERT																
Saha et al.	CV	GPT-4	0.97	0.06	0.03	0.25	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.94	0.03	0.02	0.25
		BERT					1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.93	0.05	0.03	0.50
		RoBERTa					1.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.94	0.10	0.05	0.75
		DistilBERT																
	HD	GPT-4	0.88	0.35	0.28	0.45	0.91	0.17	0.24	0.13	0.92	0.41	0.45	0.38	0.90	0.21	0.24	0.19
		BERT					0.92	0.24	0.35	0.19	0.92	0.47	0.43	0.52	0.91	0.20	0.26	0.16
		RoBERTa					0.91	0.26	0.32	0.22	0.91	0.40	0.38	0.42	0.91	0.28	0.33	0.25
		DistilBERT																

Table A4: Complete task-by-task classification performance results. Ac., Pr., and Re. refer to accuracy, precision, and recall, respectively.

Study	Task	Hyperparameters
Card et al.	Classify immigration speeches	learning rate (5e-05), batch size (8), epochs (4)
	Classify pro-immigration speeches	learning rate (5e-05), batch size (16), epochs (6)
	Classify anti-immigration speeches	learning rate (5e-05), batch size (8), epochs (6)
	Classify neutral immigration speeches	learning rate (5e-05), batch size (8), epochs (4)
Hopkins et al.	Classify race/ethnicity	learning rate (5e-05), batch size (8), epochs (4)
	Classify gender	learning rate (5e-05), batch size (8), epochs (6)
	Classify political groups	learning rate (5e-05), batch size (16), epochs (6)
	Classify religious groups	learning rate (5e-05), batch size (8), epochs (6)
Müller	Classify past	learning rate (5e-05), batch size (8), epochs (4)
	Classify present	learning rate (5e-05), batch size (8), epochs (4)
	Classify future	learning rate (2e-05), batch size (8), epochs (6)
Peng et al.	Classify criticism	learning rate (5e-05), batch size (8), epochs (6)
Saha et al.	Classify fear speech	learning rate (5e-05), batch size (8), epochs (6)
	Classify hate speech	learning rate (5e-05), batch size (8), epochs (4)

Table A5: Hyperparameter settings per task.

	BERT- base	RoBERTa- base	DistilBERT	XLNet- base	Mistral- 7B
# parameters	110m	125m	66m	110m	7b
# attention heads	12	12	12	12	32
Hidden dim.	768	768	768	768	4096
Feedforward dim.	3072	3072	3072	3072	14336

Table A6: Model architectures and additional hyperparameters.

Data set and task	BERT F1 score (train- ing obs w/o noise)	BERT F1 score (train- ing obs w/ noise)	Difference
Hopkins (AJPS): Political	0.340	0.344	-0.004
Hopkins (AJPS): religion	0.609	0.609	0.000
Hopkins (AJPS): gender	0.716	0.684	0.032
Hopkins (AJPS): race	0.635	0.640	-0.005
Muller (JOP): future	0.851	0.851	0.000
Muller (JOP): past	0.791	0.755	0.036
Muller (JOP): present	0.606	0.601	0.005
Card (PNAS): cat_imm	0.832	0.815	0.017
Card (PNAS): cat_anti	0.565	0.573	-0.008
Card (PNAS): cat_neutral	0.385	0.428	-0.043
Card (PNAS): cat_pro	0.448	0.436	0.012
Peng (PNAS)	0.431	0.444	-0.013
Saha (PNAS): CV	0.031	0.059	-0.028
Saha (PNAS): HD	0.210	0.276	-0.066

Table A7: Comparing BERT F1 score for models fine-tuned with and without noise

Clustering Document Parts: Detecting and Characterizing Influence Campaigns from Documents

Zhengxiang Wang

Department of Linguistics and IACS
Stony Brook University
zhengxiang.wang@stonybrook.edu

Owen Rambow

Department of Linguistics and IACS
Stony Brook University
owen.rambow@stonybrook.edu

Abstract

We propose a novel clustering pipeline to detect and characterize influence campaigns from documents. This approach clusters parts of document, detects clusters that likely reflect an influence campaign, and then identifies documents linked to an influence campaign via their association with the high-influence clusters. Our approach outperforms both the direct document-level classification and the direct document-level clustering approach in predicting if a document is part of an influence campaign. We propose various novel techniques to enhance our pipeline, including using an existing event factuality prediction system to obtain document parts, and aggregating multiple clustering experiments to improve the performance of both cluster and document classification. Classifying documents after clustering not only accurately extracts the parts of the documents that are relevant to influence campaigns, but also captures influence campaigns as a coordinated and holistic phenomenon. Our approach makes possible more fine-grained and interpretable characterizations of influence campaigns from documents.

1 Introduction

Inspired by [Martin et al. \(2023\)](#) and [Luceri et al. \(2023\)](#), we define an influence campaign as a *coordinated and strategic effort* to shape and manipulate the perceptions of a target audience about certain things or issues over a period of time. It can be organized by an individual, organization, or government for various purposes, such as promoting a specific public image, product, policy, or political narrative. It can be carried out through various channels, including traditional media and online platforms. Consequently, detecting an influence campaign requires *holistic evaluations* and the use of multiple indicators, such as the social network ([des Mesnards and Zaman, 2018](#)), that point to a collective effort with a shared motive that aims to

impact public opinions in a certain way. Accurate and reliable detection typically involves extensive manual verification by domain experts, taking into account both textual and non-textual information ([Martin et al., 2023](#)).

In the context of NLP, detecting influence campaigns typically means predicting if an input document is part of an influence campaign ([Luceri et al., 2023](#)), i.e., a *binary classification task*. However, this is a different task from capturing the phenomenon of influence campaigns, which naturally is a *clustering problem*, i.e., grouping a collection of documents that reflect an influence campaign.

In practice, the classification task is difficult if not doomed, because by definition, an influence campaign cannot possibly be inferred from a single document. Consider the examples in [Table 1](#), where the texts in the “Positive” column reflect an influence campaign linked to the Ukraine bioweapons conspiracy theory¹ and the texts in the “Negative” column do not. The only thing that connects the positive texts and distinguishes them from the negative texts is the *shared theme/belief*, expressed by *some parts* of each document (short or long), that there exist US biolabs in Ukraine for the purpose of developing bioweapons. Arguably, any text classifier trained on some specific influence campaign datasets will at best be reduced to detecting some key words expressing the themes of the influence campaigns in the training data; such a classifier will have brittle generalization capacity. Moreover, having a binary classification decision about whether a document reflects an influence campaign neither tells us *how* the document reflects an influence campaign, nor does it reveal *what the influence campaign is about*. In contrast, if we have a cluster of documents relevant for an influence campaign clustered together, such as those in [Table 1](#), it not only makes it possible to charac-

¹https://en.wikipedia.org/wiki/Ukraine_bioweapons_conspiracy_theory.

Media	Positive	Negative
Twitter	...Putin cleans up the bioweapons labs installed by the deep state... (44 tks)	...RT @EmmanuelMacron France strongly condemns Russia’s decision to wage war on Ukraine... (19 tks)
Forum	...a secret NATO laboratory for biological weapons...Biological weapons tests were carried out in the laboratories of this facility... (638 tks)	...[NATO] has blocked Ukraine’s plan to enter...Item 3: Ukraine was a pawn that the Westerners deliberately sacrificed to strengthen NATO... (703 tks)
News	...a NATO secret biological laboratory with biological weapons...The biological laboratory under the Azovstal plant in Marioupol in the so-called PIT-404 facility was built ...In the laboratories of the facility, tests were carried out to create biological weapons... (1497 tks)	...Russia’s demand for neutrality...But NATO members said that Ukraine’s membership was at best a distant option... [The leader of the Ukrainian separatist region of Lugansk said he could hold a referendum on integration into Russia,] a decision immediately criticized by Kiev...(1152 tks)

Table 1: Two clusters of document parts in three media forms that reflect (positive) or do not reflect (negative) an influence campaign related to the Ukraine bioweapons conspiracy theory. The number of tokens in each document is indicated in the parenthesis following each text. We use “...” to highlight and separate the parts of document that are inside the two clusters. The document parts inside the positive cluster, called high-influence cluster in Sec 3, present evidence of an influence campaign and the related documents are thus likely to be linked to an influence campaign. Here the document parts are beliefs of the author (see Sec 5) with the target words highlighted in bold.

terize the theme of the influence campaign, but it also makes it much more straightforward to understand why each document is part of an influence campaign: it is because the document, along with other documents in the cluster, contains certain document parts that express an orchestrated theme.

In this paper, we propose a novel text-only *clustering-based* pipeline to help detect and characterize influence campaigns from documents. Unlike the typical document-level classification approach discussed above, the pipeline *predicts influence campaigns directly on the cluster level*, i.e., it predicts whether a cluster of document parts present an influence campaign (a high-influence cluster). From there, the pipeline further predicts whether any document associated with a high-influence cluster is part of the influence campaign *via a dynamic projection procedure*. As a result, **our pipeline is capable of handling the two aspects of the influence campaign detection task: capturing influence campaigns as a holistic phenomenon, and predicting documents that are part of an influence campaign**. Since influence campaigns are captured by clusters of document parts and the documents predicted to engage in an influence campaign are projected from these clusters, our pipeline enables *fine-grained and interpretable characterizations* of influence campaigns

from documents. The specific contributions of this paper are as follows.

- We introduce a novel clustering pipeline that detects influence campaigns on both cluster and document levels. This approach significantly outperforms the direct document-level classification approach and the document-level clustering approach. We do not use lexical features in any of our experiments so as not to overfit to the dataset we use.
- We propose a new approach to the classification of documents based on clustering parts of the document. We show that this approach outperforms clustering documents for our task. This approach makes possible fine-grained and interpretable characterizations of *what parts of a document lead to the classification of the document*.
- We present the very first study to use multi-word text spans expressing certain belief of an entity about the factuality of an event in the input text to extract document parts. We show that for the influence campaign detection task, clustering these text spans can improve the detection performance of influence campaigns from documents, compared to simply clustering sentences.

- We show that instead of optimizing for a clustering algorithm and its parameters, using an aggregation of algorithms and parameters performs better in our classification task, and provides for more stable results.

This paper is structured as follows. We review related works in Sec 2 and motivate and explain the algorithmic idea underlying our novel clustering pipeline in Sec 3. We describe the influence campaign detection dataset we test our pipeline on in Sec 4 and the experiments in Sec 5. The results are discussed in Sec 6. The paper concludes in Sec 7. We release our code at <https://github.com/jaaack-wang/detect-influence-campaigns>.

2 Related Work

There have been very few studies in the existing literature that approach influence campaigns in the general sense as we define it in Sec 1. The influence campaigns studied in most previous research (des Mesnards and Zaman, 2018; Luceri et al., 2023; Martin et al., 2023) are political influence campaigns, or some closely related political influence operations that may be an influence campaign, such as the spreading of mis/dis-information (Ferrara, 2017; Álvaro Figueira and Oliveira, 2017; Rubin, 2017; Addawood et al., 2019; Barrón-Cedeño et al., 2019; Nogara et al., 2022; Sakketou et al., 2022; Malik et al., 2023).

The most common detection method relevant for influence campaigns is the detection of bots in social networks (Davis et al., 2016; Badawy et al., 2018; des Mesnards and Zaman, 2018; Himelein-Wachowiak et al., 2021; Hajli et al., 2022; Rossetti and Zaman, 2022). For text-based influence campaign detection, various NLP methods have been explored. For example, a recent study leverages LLMs (Luceri et al., 2023) to predict if a tweet is part of an known influence campaign. Other studies relevant for influence campaigns utilize various sources of linguistic features (e.g., lexicon counts, ngrams, word embedding) to train or fine-tune different models (e.g., BERT, graph neural network, decision tree) with a goal to detect propagandistic, deceptive, or misleading information (Addawood et al., 2019; Barrón-Cedeño et al., 2019; Sakketou et al., 2022; Malik et al., 2023). To the best of our knowledge, we are not aware of any study that aims to detect influence campaigns on the cluster level.

Several twitter datasets have been used by recent studies on detecting influence campaigns (des

Mesnards and Zaman, 2018; Luceri et al., 2023), such as the 2016 US election dataset (Littman et al., 2016), data from Twitter’s Information Operations archive², and Russian troll accounts for 2016 US election released by the U.S. Congress (Addawood et al., 2019). There are also other relevant datasets in other media forms, such as FACTOID (Sakketou et al., 2022) collected from Reddit and Propy (Barrón-Cedeño et al., 2019) collected from news articles. We note that for all of these datasets that come with labels, the labels are typically created on the basis of some simplistic association or assumption. For example, tweets are assumed to be linked to an influence campaign if they come from Russian troll accounts (Luceri et al., 2023). To the best of our knowledge, we do not know of any publicly available influence campaign datasets that contain more than one media type.

3 Pipeline: The Algorithmic Idea

Given the coordinated nature of influence campaigns, an influence campaign can be thought of as a cluster of documents that spread a certain theme aimed to influence the target audience. Our pipeline follows exactly this intuition and transforms the task of influence campaign detection into one that detects clusters that are highly likely to reflect an influence campaign (i.e., high-influence clusters). Then the next step naturally becomes how to accurately select documents (i.e., high-influence documents) associated with the high-influence clusters that reflect an influence campaign, assuming the clusters may contain some noise or false positives.

More concretely, our pipeline consists of the following four steps.

Determining document parts In a pre-processing step, we start out by extracting parts from a document. In this paper, we experiment with three types of document parts: the multi-word text spans that represent what the author expresses certain belief in (see Sec 5.3); sentences; and the whole document.

Clustering parts of documents Given a set of documents, the pipeline clusters the document parts. Clustering parts of documents not only creates a complex connection network among documents via their semantically related parts, but also presents a general and effective workaround for

²E.g., https://blog.twitter.com/en_us/topics/company/2020/2020-election-changes

long document information retrieval using unsupervised clustering algorithms (Mekontchou et al., 2023).

Classifying high-influence clusters At training time, the pipeline takes as inputs a collection of documents, each of which is annotated with a binary label: the document is or is not part of an influence campaign. The fact that our pipeline requires annotated documents during training highlights that it is a supervised approach. The concept of high-influence clusters is defined by a user-given threshold α , denoting the minimum percentage of document parts in a cluster linked to documents from an influence campaign for the cluster to count as a high-influence cluster.

Intuitively, α should be set far greater than 0.5 to align with the heuristic that a high-influence cluster should be dominated by document parts from documents that are part of an influence campaign. The assumption here is that parts of documents with a link to an influence campaign are *unlikely* to be clustered together unless they are related to some aspect or surface theme of the influence campaign. Since high-influence clusters may be rare or even absent for a given clustering experiment for which the majority of documents are “innocent” and do not reflect an influence campaign, we propose doing multiple clustering experiments and aggregating the resulting clusters together as a way to generate more data to train a classifier for high-influence clusters. This approach is a novel *data augmentation* technique for cluster-level classification. In this paper, we set $\alpha = 0.95$. We set $\alpha \neq 1$ as a trade of the precision and recall for discovering high-influence clusters, since allowing a small error term $1 - \alpha$ ($\alpha \neq 1$) in the definition of high-influence clusters facilitates discovery of more high-influence clusters, i.e., a great improvement in recall at a small cost of precision, ultimately leading to a better F1. Note that α is only set and used at the training time.

At inference time, the pipeline deploys the pre-trained classifier to detect high-influence clusters by predicting the likelihood of a cluster being a high-influence cluster.

Classifying high-influence documents High-influence documents are documents with connections to high-influence clusters, meaning at least some of their parts occur in at least one high-influence clusters. Formally, we set a thresh-

old β , a ratio for the number of high-influence clusters, which denotes the minimum number of times parts of a document that must occur in any high-influence clusters to qualify the document as a high-influence document. The threshold β is used to control the number of false positives (i.e., documents with no link to an influence campaign occurring in high-influence clusters) introduced by the threshold α set in the previous step. We wish to come up with a module in our current system that predicts an optimal threshold β in the future.

In summary, we have introduced the notion of “high-influence clusters” based on parts of documents and the notion of “high-influence documents” based on their association with high-influence clusters. We propose two thresholds (α and β) to regulate the number of false-positive high-influence documents our system may end up selecting from high-influence clusters. The threshold α is used for training a high-influence cluster classifier, whereas the threshold β is used at the time of classifying high-influence documents. We also propose the aggregation of clustering experiments, instead of hard fine-tuning for an optimal clustering experiment, to reliably enhance model performance.

In what follows, we show that our approach can easily and significantly outperform direct document-level classification, in an apples-to-apples comparison, when it comes to detecting influence campaigns from documents.

4 Data

We use data collected during a large research program, DARPA INCAS project³. We expect the data to be made public after the end of the research program. We use this dataset as we are not aware of any other datasets that have expert-verified annotations indicating if a collection of documents contain influence campaigns.

The data contains four piles of online posts published during January 31 to June 30, 2022. Each pile is a collection of documents in six media forms, namely, Twitter, Forum, News, Blog, Reddit, and Other. Two of the four piles contain documents that engage in an influence campaign that spreads disinformation related to Ukraine bioweapons conspiracy theory (see Table 1), whereas the other two contain Ukraine-related documents with no links

³<https://www.darpa.mil/program/influence-campaign-awareness-and-sensemaking>

to any known influence campaigns. Lexical search was used to facilitate the collection of the data. Over 99% of the documents that participate in the bioweapons influence campaign use words like “biolab” and “biological weapons”, but slightly less than 3% of the documents unrelated to the campaign mention these terms. That means that any content-based text classifier, whether rule-based or neural, will overfit this dataset by capturing the related keywords. We avoid training such classifiers, as (1) we are more interested in developing a *potentially general* approach that can both detect and characterize influence campaigns from documents; and (2) content-based text classification for the influence campaign detection task arguably cannot be a general approach nor can it make predictions at beyond the document level to capture the phenomenon of influence campaign.

The majority of these documents are written in French, typically accompanied by an English translation, and the rest are in English. We choose to work on the translated French portion of the data. This portion has over 8 times more documents than the English subcorpus, but with a significantly smaller portion of documents linked to an influence campaign, less than 8%. We believe this represents a more realistic and challenging setting for detecting influence campaigns from documents.

Given the overall small size of the dataset, we randomly split the French data into two parts, the train and test sets, but the train set can be further split for training and validation where needed. We split at the document level, with a ratio of 80/20, as shown in Table 2. Appendix A provides further details about the distribution of media forms and average document length in the data.

5 Experiments

5.1 Task description

As argued in Sec 1, the real challenge for detecting influence campaigns is how to capture the phenomenon of influence campaigns. Making the detection of influence campaigns a binary classification task, i.e., to predict whether a text or a document is part of an influence campaign, is not only less realistic but also probably doomed; such a classification approach cannot survive the constantly shifting and evolving nature of influence campaigns as a dynamic social phenomenon.

Nevertheless, to comprehensively evaluate our new pipeline requires a large-scale dataset anno-

	Train	Test
# Docs	5334 (416; 7.8%)	1333 (56; 4.2%)
# Sents	72,330 (15,394; 21.3%)	14,370 (2,182; 15.2%)
# Targets_{ALL}	270,818 (61,652; 22.8%)	50,781 (8,531; 16.8%)
# Targets_{AT}	155,238 (34,703; 22.4%)	29,793 (4,905; 16.5%)

Table 2: Statistics of the train and test sets (disjoint). Numbers inside parentheses show the number and percentage of documents or document parts linked to an influence campaign. Targets_{AT}: targets the Author herself believes to be True. Targets_{ALL}: all belief targets, regardless of the belief holder and the commitment level.

tated on the document collection level, indicating if a collection of documents presents an influence campaign. Since the dataset described in Sec 4 is the only dataset in this regard, this comprehensive evaluation cannot be possible.

Instead, we have to resort to the detection of influence campaigns as a binary classification task at the document level. This allows us to quantitatively compare our clustering approach with the existing classification approach and validate its potential as a general method to detect and characterize influence campaigns from documents.

As a classification task, the objective is to accurately identify as many documents as possible that are linked to the known bioweapons influence campaign. Since such documents in our dataset are rare, we use precision, recall, and binary F1 to measure the classification performance of the examined approaches, which also helps us to understand the types of errors these approaches make.

5.2 Baselines

We train two direct document-level classifiers (**Direct-document**), using fully connected feed-forward neural networks (FNN) and XGBoost algorithm (Chen and Guestrin, 2016), as the baselines to compare with our approach. XGBoost is an optimized gradient boosting (Friedman, 2001) system using tree ensembles that achieve state-of-the-art results on many real-world machine learning challenges. We refrain from using any word-embedding-based or content-based machine learning models to prevent models from learning from general lexical data, which precludes the use of models such as LLMs, BERT, LSTM, and so on.

We use 95 general linguistic features, extracted by an open-sourced corpus linguistics tool (Wang, 2021), to train the models. These features are mostly based on the work of Biber (1988, 2006) and have been developed over decades to suit *general text analysis*. We also add “number of words” as a feature to factor in document length, particularly for short documents (say, a tweet), which may not see the occurrence of many features at all due to length limitation. The extracted features are mostly normalized frequency counts. More details about the model parameters and these features are given in Appendix B.

In addition, we apply our pipeline on the document level (**Document-level**), namely, clustering the entire documents, as an additional baseline to emphasize the importance of clustering document parts. The experimental setup for document-level clustering is identical to the setup for our approach based on document parts.

5.3 Our approach

Obtaining document parts We break down a document into parts in three ways. (1) We use the sentences of the document as its parts (**Sentence-level**). We use the default sentence segmentation algorithm from spaCy V3.5.3. (Honnibal et al., 2020) (2) We also experiment with the state-of-the-art event factuality prediction system (Murzaku et al., 2023) to extract from each sentence of a document (*source*, *target*, *factuality label*) triplets. Here, *source* refers to the belief holder, *target* is a head word denoting an event, and *factuality label* describes the extent to which the source believes that the event has happened, is happening, or will happen. The source can either be the author herself, or somebody else according to the author. The factuality label has five possible values, ranging from committed belief (certain that true) to committed disbelief (certain that false), with possible belief, unknown belief, and possible disbelief in between. We use a head-to-span algorithm to extract a multi-word text span, of which the target is the syntactic head, as the representation of the identified event to be used as the extracted document parts. For **Target_{ALL}-level**, we use all target spans extracted by the belief system. (3) We use the same event factuality prediction system but we retain only the events believed by the author (**Target_{AT}-level**), a subset of all the events identified by the event factuality prediction system, to see if document parts

to which the author holds a belief will lead to a better result using our approach. The examples in Table 1 are events believed by the author, where the target words are highlighted in bold. The number of sentences and targets in the train and test sets is listed in Table 2.

Clustering We use S-BERT (Reimers and Gurevych, 2019) to embed document parts and then employ two clustering algorithms, i.e., KMEANS (MacQueen, 1967) and HDBSCAN (Campello et al., 2015), to cluster the embedded document parts. HDBSCAN is a hierarchical extension of DBSCAN (Ester et al., 1996) with various optimization methods implemented (Campello et al., 2015). Due to the curse of dimensionality (Bellman, 1961), HDBSCAN does not easily produce clusters without an embedding reduction algorithm in place. We use the state-of-the-art UMAP algorithm (McInnes et al., 2018) for this purpose.

Classifying high-influence clusters To contrast the baselines with our approach, we use the same two classification algorithms (FNN and XGBoost) to classify high-influence clusters. In addition to the 95 general linguistic features, there are 7 cluster-level features that are specific to our current pipeline: top-10 uni-gram text frequency, top-10 bi-gram text frequency, top-10 tri-gram text frequency, weighted n-grams text frequency, average cosine similarity between all pairs of document parts in the cluster, percentage of unique documents, and cluster size. Top-10 n-grams text frequency is the average ratio of texts containing the top-10 n-grams, whereas the weighted n-grams text frequency is the weighted sum of the aforementioned top-10 n-grams text frequencies, where the weights are simply given as $\frac{n}{\sum_{n'=1}^3 n'}$ for $n \in \{1, 2, 3\}$. Average cosine similarity (ACS) is the average cosine similarity of all unique text pairs in a cluster:

$$ACS = \frac{\sum_i^m \sum_{j \neq i}^m \cos_sim(text_i, text_j)}{m(m-1)}$$

These 5 features are designed as “hard” (ngrams) and “soft” (ACS) measurements of topical and thematic coherence of a cluster, which are relational and independent of the specific lexical choices used inside the cluster. Percentage of unique documents and cluster size are just basic attributes of a cluster. The percentage of unique documents is calculated by dividing the number of documents whose parts

occur in the cluster by the number of document parts in the cluster.

Cluster aggregation We run a total of 135 clustering experiments by varying related parameters for the two clustering algorithms we use, detailed in the Appendix B. We use all resulting clusters both for training the high-influence cluster classifier, as well as for selecting high-influence documents from high-influence clusters.

Classifying high-influence documents For a single clustering experiment, we simply classify any document, whose parts occur in at least one high-influence cluster, as a high-influence document. This low threshold is to ensure that short documents would not be excluded from being identified as high-influence documents, since they may only be segmented into one part and cannot have more than one association with high-influence clusters.

As mentioned, we propose using all high-influence clusters from multiple clustering experiments to expand the search for high-influence documents. However, as a result of this aggregation, the chance of misidentifying a high-influence document based on a single association with a high-influence cluster increases, since false positives in high-influence clusters also accumulate with aggregation. To regulate the false positive rate, we set $\beta = 0.2$ (see Sec 3 for definition).

Evaluation Since a clustering configuration does not necessarily produce high-influence clusters and in practice we can always try to find one that does, we evaluate the average performance of our approach on various clustering setups that produce high-influence clusters predicted by our pipeline. We choose a wide range of clustering configurations (see Appendix B) so as to avoid hard fine-tuning our approach. We run the two classification algorithms (FNN and XGBoost) for five times with varying parameters to compare the average performance of our approach against the baseline approaches on our dataset.

6 Results

6.1 Main findings

The document classification approach versus ours Table 3 shows the main results of the experiments. As expected, our approach significantly outperforms the direct document-level classification approach.

Clustering documents versus document parts

Clustering document parts clearly outperforms clustering documents by a significant margin. When FNN is used to classify high-influence clusters, clustering documents barely works at all.

FNN versus XGBoost Our models that use XGBoost to classify clusters achieve overall high precision, regardless of aggregation. Those which use FNN suffer from low precision without aggregation. This means that high-influence clusters predicted by XGBoost contain much fewer false positives, i.e., associated documents with no link to an influence campaign, than those predicted by FNN. This makes XGBoost a better choice for cluster prediction for the current paper.

Document parts There is value in clustering belief targets. They are multi-word text spans within a sentence that carry a factuality label and involve a belief source. Compared to full sentences, they are more information-dense. We find that when FNN is used to classify clusters, clustering belief targets the author holds to be true (Target_{AT}-level) leads to the best performance, independent of the use of aggregation. When XGBoost is used, clustering all belief targets outperforms clustering sentences by 18% absolute in F1 without aggregation. These results show the potential of extracting belief targets for a better detection of influence campaigns, which intuitively make sense because influence campaigns are all about spreading a certain belief of the influencers.

Cluster aggregation From Table 3, we see that cluster aggregation helps in every experiment (leaving aside document-level clustering using FNN, which performs at near-0 levels). Most of these improvements are statistically significant, given the standard deviations shown. However, for FNN using Target_{AT}-level there is no significant difference, and for XGBoost using Target_{ALL}-level there is no significant difference. We have no explanation for these exceptions for now. In general, cluster aggregation helps in two ways. When our models have very low precision (using FNN), the current aggregation setup rules out many false positives resulting from misclassified high-influence clusters, which greatly improves precision. On the other hand, when precision is decent (in the case of XGBoost), aggregation can serve to help increase the range of relevant documents associated with high-influence clusters, which lead to a better recall in

	FNN			XGBoost		
	Precision	Recall	F1	Precision	Recall	F1
Direct-document	20.2 \pm 2.2	18.9 \pm 14.9	17.1 \pm 8.6	77.3 \pm 9.3	37.9 \pm 7.9	50.7 \pm 9.1
Document-level (mean)	0.3 \pm 0.1	0.7 \pm 0.1	0.4 \pm 0.1	90.7 \pm 2.5	25.4 \pm 3.3	38.2 \pm 4.6
+ Aggregation	0.0 \pm 0.0	0.0 \pm 0.0	0.0 \pm 0.0	94.1 \pm 0.7	28.6 \pm 3.1	43.8 \pm 3.8
Sentence-level (mean)	28.3 \pm 4.1	44.1 \pm 4.7	32.8 \pm 4.2	69.4 \pm 10.9	50.4 \pm 2.7	56.7 \pm 4.1
+ Aggregation	74.5 \pm 16.4	43.2 \pm 4.1	54.3 \pm 7.7	86.5 \pm 1.8	70.7 \pm 2.4	77.8 \pm 2.0
Target _{ALL} -level (mean)	25.4 \pm 6.7	35.2 \pm 8.5	27.0 \pm 6.9	78.2 \pm 3.7	73.8 \pm 2.4	75.3 \pm 1.3
+ Aggregation	72.5 \pm 4.5	40.0 \pm 5.7	51.5 \pm 5.7	81.1 \pm 3.5	71.1 \pm 7.3	75.5 \pm 3.8
Target _{AT} -level (mean)	60.7 \pm 7.1	66.8 \pm 10.5	62.4 \pm 8.5	63.5 \pm 2.2	49.5 \pm 2.8	54.8 \pm 2.1
+ Aggregation	64.8 \pm 4.6	61.8 \pm 8.6	63.1 \pm 6.0	80.2 \pm 3.5	71.4 \pm 1.8	75.5 \pm 0.9

Table 3: Average model test set performance plus standard deviation (in %) under different training conditions for five runs. For the clustering approach, when there is no aggregation of high-influence clusters, the performance is the average performance across different clustering experiments averaged over five runs. Note that the two classification algorithms (FNN and XGBoost) are only used to classify clusters in our approach.

Media (pos/neg)	Direct-Doc		Doc-level		Our approach	
	FN	FP	FN	FP	FN	FP
Twitter (11/686)	11.0	0	8.8	0	3.0	1.0
Forum (7/136)	5.4	0	6.0	0	1.0	0
News (24/280)	11.4	4.6	17.0	0	10.4	4.2
Blog (13/62)	6.0	1.4	7.8	1.0	2.0	1.0
Reddit (0/91)	NA	0	NA	0	NA	0
Other (1/22)	1.0	0	0.4	0	0	0
Total (56/1277)	34.8	6.0	40.0	1.0	16.4	6.2

Table 4: Average test set error counts of the three best models from the two baseline approaches plus ours across the six media types over the five runs. The first column indicate the media type along with the numbers of documents that reflect (“pos”) or do not reflect (“neg”) an influence campaign. FN: False Negative. FP: False Positive. NA is due to zero positives in Reddit. The best model configurations for the three approaches all use XGBoost and aggregation where applicable. Our best system is on the sentence level per Table 3.

most cases.

6.2 Error analysis

Table 4 reports the average number of errors made by the two baseline approaches and our approach on the test set, over the five runs. The error counts are broken down according to media types.

Direct-document The direct document-level classification approach fails to recognize all the documents from Twitter and Other that reflect the bioweapons influence campaign. It also misses 5.4 out of 7 documents from Forum on average across five runs, with slightly less than 50% false negative rate for documents from News and Blog. Given the average document length for these five media types (see Table 5), it is clear that this classification

approach works poorly on identifying short documents linked to an influence campaign when training on documents with a wide length range. This is probably because the model learns some discriminative features from long documents, which may not be observed in short documents. Conversely, a similar issue may also occur the other way around, suppose the model trains on short documents. This may be one of the inherent limitations of the direct document-level classification approach, even when the models are trained and deployed for predicting an known influence campaign.

Document-level Directly clustering documents allows the model to recognize influence campaigns in documents of different genres and length, which is an advantage compared to the direct document-level classification approach. However, clustering documents also makes it hard for the model to efficiently identify information in the documents related to influence campaigns, which may exist only in parts of document. This results in the high number of false negatives. That said, our current pipeline setup helps this document-level clustering approach to make accurate positive predictions, given the lowest number of false positives.

Our approach Clearly, clustering document parts helps overcome the limitations faced by the two baseline approaches, since the model recognizes influence campaigns in documents irrespective of the genre and have less than half false negatives, compared to the other two approaches. We identify 19 documents (15 FNs and 4 FPs) where the models across the five runs consistently mis-

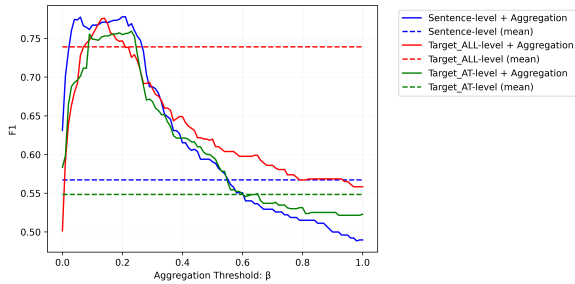


Figure 1: Aggregation versus no Aggregation with XGBoost as the high-influence cluster classifier. Results are averaged over the five runs.

classify. By using the keyword “bio”, we identify 3 of the 19 documents that may be mislabelled. For the remaining 16 misclassified documents, we hypothesize that the errors are mainly caused by two reasons. First, sentences do not necessarily reflect the theme of the document, which, for example, may make our model confuse documents exposing an influence campaign with one that spreads it. Second, none of the other techniques (e.g., SBERT, clustering) used in our pipeline are free of errors, which can propagate and ultimately lead to a wrong classification decision. We wish to improve our pipeline along these two directions in the future.

6.3 Threshold for models with aggregation

Concerns may arise over our use of a less justified threshold β to select documents from high-influence clusters as high-influence documents. This threshold is a ratio of the number of high-influence clusters available for aggregation. As explained in Sec 3, this ratio helps prevent the false negatives in each high-influence cluster from accumulating uncontrolled, as a result of aggregation.

Nevertheless, as shown in Fig 1 (also see Fig 2 from Appendix C for FNN), the classification F1 with aggregation is almost always better than without aggregation for most models. Unsurprisingly, the performance curve shows an upside down U-shape, a trade-off between precision and recall as we vary β . Setting $\beta = 0.2$ is a conservative choice, which does not lead to the optimal performance. In the future, we would like to explore an automatic way of finding the optimal value for β .

7 Conclusion

We have presented a new approach to finding influence campaigns, which relies on four core features: (1) we cluster parts of documents; (2) we classify

clusters of parts of documents using non-lexical features; (3) we relate the classification result back to documents; (4) we use cluster aggregation, the use of many clustering runs over the same dataset, to augment training data for the cluster classifier. The resulting classification of the documents does not only show a predicted label for the document (part of influence campaign or not), but it also shows which parts of the document are responsible for this classification. We believe that our general approach can profit other document classification tasks, including detecting scientific influence in published papers, or themes in literature.

There are several avenues for possible future work and we list three below. (1) **Datasets**. Given the increasing importance of detecting influence campaigns, we hope there will be more datasets annotated on the document collection level for an influence campaign. (2) **Incorporating non-textual information**. Our current pipeline is a text-only system. Leveraging non-textual information, such as social interactions and the authors’ past activities, may help us create a more complicated and comprehensive system (e.g., using graph neural network) that enhances the accurate and reliable detection of influence campaigns. However, such work cannot be possible without good datasets. (3) **Automatic characterization of influence campaigns**. Our work captures influence campaigns by the high-influence clusters, which may contain a large number of semantically related document parts, possibly with noise. To fully make sense of these clusters, we need to have some automatic ways of characterizing them in a fine-grained and interpretable way aligned with the downstream needs. Our preliminary experiments show that LLMs may be a potential option.

Acknowledgement

We thank three anonymous reviewers from the 6th NLP+CSS Workshop for the constructive and helpful comments. This material is based on work supported by the Defense Advanced Research Projects Agency (DARPA) under Contracts No. HR01121C0186, No. HR001120C0037, and PR No. HR0011154158. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA. Rambow gratefully acknowledges support from the Institute for Advanced Computational Science at Stony Brook

University.

Limitations

This study serves as a preliminary evaluation and validation of the new paradigm we propose for influence campaign detection. Given the lack of data and constrained by time, we have not been able to show that the approach also works on an entirely unseen dataset (though of course we tested it on unseen documents in our dataset).

- We define influence campaigns in a very general sense, but our approach is only tested on data relating to political influence campaigns. We need to test our approach on other non-political influence campaign datasets.
- We cannot release the dataset we used to train and test our pipeline due to the funding agency’s restrictions. We hope once the current program is finalized, the dataset will be released so that our study can be reproduced.
- We did not spend a large amount of time attempting to improve the direct-document approach. We cannot guarantee that with a different set of (non-lexical) features and well-tuned parameters, a direct document-level classifier may not outperform our approach.

Ethical Concerns

Working with social media often brings privacy concerns. The data we are working with has already been anonymized. For example, Twitter handles have been replaced by random designators. Furthermore, in our work, we do not use any part of the information about the author, we only use the text.

References

- Aseel Addawood, Adam Badawy, Kristina Lerman, and Emilio Ferrara. 2019. Linguistic cues to deception: Identifying political trolls on social media. In *International Conference on Web and Social Media*.
- Adam Badawy, Emilio Ferrara, and Kristina Lerman. 2018. Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 258–265.
- Alberto Barrón-Cedeño, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. 2019. Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management*, 56(5):1849–1864.
- Richard E. Bellman. 1961. *Adaptive control processes: a guided tour*. Princeton University Press.
- Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press.
- Douglas Biber. 2006. *University Language: A Corpus-Based Study of Spoken and Written Registers*. John Benjamins Publishing.
- Ricardo J. G. B. Campello, Davoud Moulavi, Arthur Zimek, and Jörg Sander. 2015. Hierarchical density estimates for data clustering, visualization, and outlier detection. 10(1).
- Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’16*, page 785–794, New York, NY, USA. Association for Computing Machinery.
- Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web, WWW ’16 Companion*, page 273–274, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.
- Nicolas Guenon des Mesnards and Tauhid Zaman. 2018. Detecting influence campaigns in social networks using the ising model.
- Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD’96*, page 226–231. AAAI Press.
- Emilio Ferrara. 2017. Disinformation and social bot operations in the run up to the 2017 french presidential election. *First Monday*.
- Jerome H. Friedman. 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189 – 1232.
- Nick Hajli, Usman Saeed, Mina Tajvidi, and Farid Shirazi. 2022. Social bots and the spread of disinformation in social media: The challenges of artificial intelligence. *British Journal of Management*, 33(3):1238–1253.
- McKenzie Himelein-Wachowiak, Salvatore Giorgi, Amanda Devoto, Muhammad Rahman, Lyle Ungar, H. Schwartz, David Epstein, Lorenzo Leggio, and Brenda Curtis. 2021. Bots and misinformation spread on social media: A mixed scoping review with implications for covid-19 (preprint). *Journal of Medical Internet Research*, 23.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. [spaCy: Industrial-strength Natural Language Processing in Python](#).

Justin Littman, Laura Wrubel, and Daniel Kerchner. 2016. [2016 United States Presidential Election Tweet Ids](#).

Luca Luceri, Eric Boniardi, and Emilio Ferrara. 2023. [Leveraging large language models to detect influence campaigns in social media](#).

J. MacQueen. 1967. [Some methods for classification and analysis of multivariate observations](#).

Muhammad Shahid Iqbal Malik, Tahir Imran, and Jamjoom Mona Mamdouh. 2023. [How to detect propaganda from social media? exploitation of semantic and fine-tuned language models](#). *PeerJ Comput Sci*.

Diego A Martin, Jacob N Shapiro, and Julia G Ilhardt. 2023. [Introducing the online political influence efforts dataset](#). *Journal of Peace Research*, 60(5):868–876.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. [Umap: Uniform manifold approximation and projection](#). *Journal of Open Source Software*, 3(29):861.

Paul Mbate Mekontchou, Armel Fotsoh, Bernabe Batchakui, and Eddy Ella. 2023. [Information retrieval in long documents: Word clustering approach for improving semantics](#).

John Murzaku, Tyler Osborne, Amittai Aviram, and Owen Rambow. 2023. [Towards generative event factuality prediction](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 701–715, Toronto, Canada. Association for Computational Linguistics.

Gianluca Nogara, Padinjaredath Suresh Vishnuprasad, Felipe Cardoso, Omran Ayoub, Silvia Giordano, and Luca Luceri. 2022. [The disinformation dozen: An exploratory analysis of covid-19 disinformation proliferation on twitter](#). In *Proceedings of the 14th ACM Web Science Conference 2022, WebSci '22*, page 348–358, New York, NY, USA. Association for Computing Machinery.

Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Michael Rossetti and Tauhid Zaman. 2022. [Bots, disinformation, and the first impeachment of u.s. president donald trump](#). *PLOS ONE*, 18.

Victoria L. Rubin. 2017. [Deception detection and rumor debunking for social media](#).

Media	# docs (train + test)	Avg doc length
Twitter	2,795 (2,098 + 697)	26.6 \pm 13.1
Forum	887 (744 + 143)	330.7 \pm 574.4
News	2,039 (1,735 + 304)	654.8 \pm 851.9
Blog	415 (340 + 75)	945.0 \pm 1,668.1
Reddit	371 (280 + 91)	69.3 \pm 131.7
Other	160 (137 + 23)	92.0 \pm 102.3

Table 5: Number of documents for the six media types and their average number of tokens plus standard deviation (combing the train and test set).

Flora Sakketou, Joan Plepi, Riccardo Cervero, Henri Jacques Geiss, Paolo Rosso, and Lucie Flek. 2022. [FACTOID: A new dataset for identifying misinformation spreaders and political bias](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3231–3241, Marseille, France. European Language Resources Association.

Zhengxiang Wang. 2021. [A macroscopic re-examination of language and gender: A corpus-based case study in university instructor discourses](#). Master’s thesis, University of Saskatchewan.

Álvaro Figueira and Luciana Oliveira. 2017. [The current state of fake news: challenges and opportunities](#). *Procedia Computer Science*, 121:817–825. CENTERIS 2017 - International Conference on ENTERprise Information Systems / ProjMAN 2017 - International Conference on Project MANagement / HCist 2017 - International Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN/HCist 2017.

A Data

Table 5 shows the average number of documents for each one of the six media forms in our data and their average document length (including both train and test set) measured in the number of tokens. Distribution wise, the related statistics in the test set is similar.

B Experimental details

B.1 FNN details

We use a simple FNN architecture with three hidden layers whose dimensionalities are 90, 60, 30, respectively. Each layer is a fully connected layer that consists of two linear transformations with a tanh activation function in between:

$$\text{FNN_layer}(x) = \tanh(xW_1 + b_1)W_2 + b_2$$

We apply Adam optimizer with 5e-4 learning rate and 1e-5 L2 weight decay rate. We randomly

take out 20% of data from the train set to obtain a validation set, which is used for the five runs. We train the model for 500 epochs and deploy the model with the best F1 on the held out validation set to the test set for evaluation.

B.2 Linguistic features from Wang (2021)

According to Wang (2021), 2/3 of the 95 features come from Biber (2006) with 42 of them also available in Biber (1988).

These 95 features can be broken down into four categories: (1) structural features, such as mean word length, type-token ratio; (2) conversational features, such as contraction (e.g., “I am” → “I’m”); (3) sentential features, which involve features related to passive voice, tense, coordination, and WH structure etc.; (4) lexical features, including part of speech, noun sub-categories, verb sub-categories, stance-related expressions, and so on. For full details, please refer to Wang (2021).

B.3 XGBoost details

We use the default configuration of the xgboost (v1.7.3) package in Python⁴ for training the XGBoost classifiers, except for the “max_depth” parameter, which we simply make equal to the current run number (i.e., 1, 2, 3, 4, 5).

B.4 Clustering details

We use the best-performing pretrained SBERT model “all-mpnet-base-v2”⁵ to embed each text before clustering. For each clustering setup, we run the same experiments for three times to obtain small variations in the clustering results.

For KMEANS⁶, we vary the number of clusters (i.e., the k) and use the following numbers: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 150, 200, 250, 300, 500. This results in 45 ($= 15 \times 3$) different experiments.

For HDBSCAN⁷, we vary two parameters. One is the minimum cluster size, which is part of the HDBSCAN algorithm. The other is the dimensionality of the reduced SBERT embedding by UMAP, which is not part of the HDBSCAN algorithm, but essential for HDBSCAN to produce meaningful

⁴https://xgboost.readthedocs.io/en/stable/python/python_api.html

⁵https://www.sbert.net/docs/pretrained_models.html

⁶<https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>

⁷<https://hdbscan.readthedocs.io/en/latest/>

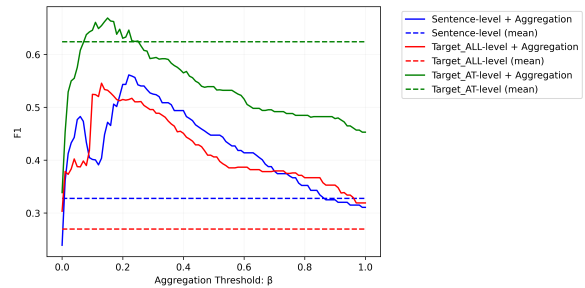


Figure 2: Aggregation versus no Aggregation with FNN as the high-influence cluster classifier. Results are averaged over the five runs.

number of clusters. We set the following minimum cluster sizes: 10, 20, 40, 80, 100, 150, 200, 300, 400, 500. The size of reduced dimensionalities are 10, 30, or 50. This results in a total of $10 \times 3 \times 3 = 90$ different experiments.

The choices of parameters are not totally random, since some of them are somehow informed by our initial experiments. But they are not cherry picked either, since we simply use a wide range of numbers to vary the related parameters, without knowing the final results.

As discussed in the paper, the main purpose for different clustering experiments is to aggregate them, either as a means of data augmentation or enhance model performance on classifying documents at the final stage of the pipeline.

C Results

Fig 2 shows the performance variation of our models at different document part levels plus aggregation, as a function of the threshold β : the minimum number of times a document must be associated with a high-influence cluster in order to qualify as high-influence document, proportional to the total number of high-influence clusters available. The mean performance of these models on clusters from each clustering experiment is shown in dashed lines, as a baseline comparison.

A First Step towards Measuring Interdisciplinary Engagement in Scientific Publications: A Case Study on NLP + CSS Research

Alexandria Leto¹ Shamik Roy^{2*} Alexander Hoyle³

Daniel E. Acuna¹ Maria Leonor Pacheco¹

¹University of Colorado Boulder ²AWS AI Labs ³University of Maryland

¹{alexandria.letto, daniel.acuna, maria.pacheco}@colorado.edu

²royshami@amazon.com ³hoyle@umd.edu

Abstract

With the rise in the prevalence of cross-disciplinary research, there is a need to develop methods to characterize its practices. Current computational methods to evaluate interdisciplinary engagement—such as affiliation diversity, keywords, and citation patterns—are insufficient to model the degree of engagement between disciplines, as well as the way in which the complementary expertise of co-authors is harnessed. In this paper, we propose an automated framework to address some of these issues on a large scale. Our framework tracks interdisciplinary citations in scientific articles and models: 1) the section and position in which they appear, and 2) the argumentative role that they play in the writing. To showcase our framework, we perform a preliminary analysis of interdisciplinary engagement in published work at the intersection of natural language processing and computational social science in the last decade.

1 Introduction

As scholarly disciplines have become increasingly specialized, there is a growing need to share knowledge across traditional interdisciplinary boundaries in order to address broader challenges (Vosskamp et al., 1986). Recognizing this trend, scholarly institutions have established interdisciplinary centers (Turner et al., 2015; Leahey and Barringer, 2020) and funding agencies are incentivizing such collaborative efforts (Hackett, 2000; Rhoten and Parker, 2024). However, it is unclear how this interdisciplinarity is reflected in the main outcome of science: scientific publications.

Existing measures of interdisciplinarity are overly simplistic and fail to capture the depth of cross-disciplinary integration in research (McCarthy and Dore, 2023). Current metrics, such

as affiliation diversity, keywords, and citation patterns (Porter and Rafols, 2009; Van Noorden, 2015; Chen et al., 2021) often overlook how co-authors' expertise is used, and keywords fail to truly reflect a paper's content. Moreover, where and how a citation is made within papers remains largely unexplored. For example, citing papers from other fields in the opening paragraph of the introduction might signal higher interdisciplinarity than doing so in a footnote of the method section. In turn, using these references to ground findings and substantiate claims signals a deeper cross-disciplinary engagement than briefly discussing the connections between two bodies of work.

To address this challenge, we propose an automated framework for characterizing and measuring the degree of meaningful cross-disciplinary engagement in scientific publications by explicitly modeling the structure (i.e., where do interdisciplinary references appear?) and content (i.e., how are these references used to make claims?) of scientific papers. To do this, we track all interdisciplinary citations in a given article. We consider a citation to be interdisciplinary if it belongs to a venue outside of the current paper discipline. Then, for each sentence containing an interdisciplinary citation, we identify its argumentation type. To define argumentation types we build on the schema proposed by Lauscher et al. (2018), which distinguishes between claims made about the author's own work, claims made about the background of the author's work, and factual statements that serve as evidence for or against a claim. Details about the implementation of this framework are provided in Sec. 3.

As a case study, we apply our framework to research papers in the interdisciplinary field of natural language processing (NLP) and computational social science (CSS). In the past decade, greater volumes of text data and more accessible methods have caused a growth in popularity of this research area (Grimmer and Stewart, 2013), including a

*Work done before joining AWS AI Labs.

dedicated workshop in leading NLP conferences*. However, like other interdisciplinary research, the synergy between NLP and CSS is often violated due to a misalignment between the research outlook, requirements, and expertise of the researchers in the corresponding fields. In recent work, [McCarthy and Dore \(2023\)](#) manually review a set of CSS papers that incorporate text analysis methods, and conclude that many of these contributions present what they call *descriptive findings*: papers that present descriptive catalogs of evidence derived from analyzing social data (e.g., word distributions found in tweets about mass shootings) but that fail to integrate these findings with relevant social science theory. They contrast these contributions with *integrative findings*, which seek to achieve synergistic methodology to meet the standards of both disciplines, furthering theory. Similarly, [Baden et al. \(2022\)](#) note that available NLP methods often fail to meet the needs of social science research, where a limited ability to incorporate theory damages methods' validity. Our work presents the first large-scale analysis of the way in which authors working on NLP+CSS have engaged with the literature in fields outside of computer science and linguistics over the past 10 years.

We make the following contributions: 1) We propose a first step towards a general computational framework to analyze interdisciplinary engagement in scientific publications. 2) We construct a comprehensive dataset of computational social science articles published in NLP venues in the last 10 years. 3) We perform a large scale analysis of interdisciplinary engagement in NLP+CSS research, and show that while interest in NLP+CSS work is growing, there is a decreasing trend in the engagement with outside disciplines in the mainstream NLP conferences. We explore these trends in the context of the main topics of interest in the NLP+CSS community and how they have shifted over time. We also show that dedicated workshops like the *NLP and Computational Social Science Workshop* attract highly interdisciplinary contributions, fulfilling their mission of providing an outlet for this type of work.

2 Related Work

Most previous studies at the intersection of NLP and the Science of Science have analyzed scientific publications by looking at their citation patterns.

*<https://aclanthology.org/venues/nlpcss/>

Some of this work has focused on the way citation behavior relates to the scientific content of articles. For example, [Jurgens et al. \(2018\)](#) studied the effect of framing contributions through citations, [Qazvinian and Radev \(2008\)](#) incorporated citation networks in document summarization, and [Cohan et al. \(2020\)](#) used citation graphs to learn scientific document embeddings. Another line of research has studied citation behavior in the NLP literature, by looking at how scientific articles are distributed across geographies ([Rungta et al., 2022](#)), or over different types of NLP papers (short, long, demo, etc.) ([Mohammad, 2020](#)).

While language-centered approaches are scarce in science of science research, there have been some prior efforts in this direction. Some notable examples are: studying cross-field jargon interpretation ([Lucy et al., 2023](#)), the influence of articles in the scientific community ([Yogatama et al., 2011](#); [McKeown et al., 2016](#); [Gerow et al., 2018](#)), the evolution of scientific topics ([Prabhakaran et al., 2016](#)), and the prevalence of different research themes ([Mendoza et al., 2022](#)).

In this paper, we look at *when* and *how* interdisciplinary citations are used in scientific articles. Previous work looking at citation context in scientific discourse has modeled the sentiment towards cited articles ([Athar and Teufel, 2012](#); [Munkhdalai et al., 2016](#)), citation intent ([Kunnath et al., 2022](#)), purpose and influence ([N. Kunnath et al., 2021](#)), and critical vs non-critical arguments ([Te et al., 2022](#)). In our study, we focus on the location and argumentation role of interdisciplinary citations.

Our work is broadly related to the argumentation mining literature ([Peldszus and Stede, 2013](#); [Lawrence and Reed, 2020](#)). We study argumentation in the context of scientific publications. While previous studies focus on identifying argumentative discourse units ([Binder et al., 2022](#)) and their relations ([Lauscher et al., 2018](#); [Gao et al., 2022](#)), we study how work coming from outside disciplines is used to make arguments in scientific articles.

3 Framework

In this section, we describe our automated framework to model the content and structure of NLP+CSS papers. Our framework is composed of three sub-tasks: 1) Identifying papers that present CSS findings and contributions, 2) For each relevant paper, identifying all cross-disciplinary citations, and 3) For each cross-disciplinary citation,

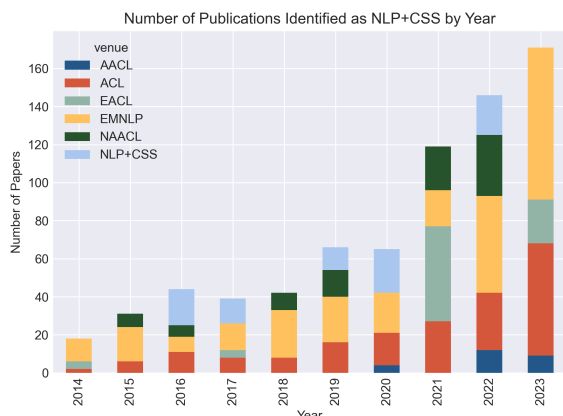


Figure 1: Resulting dataset of NLP+CSS publications over the last decade

identifying the type of argument being made.

3.1 Data Collection

To perform our analysis, we first need to construct a dataset of representative CSS articles published in NLP conferences. To do this, we first collect all long papers published in the ACL Anthology between 2014 and 2024 in all main NLP conference proceedings (ACL, EMNLP, NAACL, EACL and AACL), as well the Workshop on NLP and Computational Social Science (NLP+CSS). This results in a dataset of about 13,000 scientific papers.

Next, we need to identify which of those 13,000 papers correspond to CSS contributions. In the next section, we outline our method to achieve this.

3.2 CSS Track Identification

With the goal of building a classifier to identify CSS papers, we label a subset of about 1,800 example papers with their track to serve as training data. Namely, we collect all papers that were listed under the “Computational Social Science and Cultural Analytics” track in schedules and conference handbooks available online for the years of 2021, 2022 and 2023 and consider them as positive examples. We also add all papers published in the NLP+CSS Workshop proceedings in any given year to the set of positive examples. To generate negative examples, we follow the same procedure for each additional track (e.g., Language Generation, Machine Translation). This results in 236 positive and 1,552 negative examples.

We build a binary classifier to automatically identify papers within the dataset that fall under NLP+CSS, and are thus pertinent to our analysis. To do so, we fine-tune a pre-trained RoBERTa clas-

sifier using the abstracts of the labeled subset of data. General statistics for the resulting dataset are shown in Fig. 1, and a more detailed dataset breakdown may be found in App. A.1.

3.3 Cross-Disciplinary Citation Mapping

To study interdisciplinary engagement, we need to track all citations in papers that reference work outside of NLP, Linguistics and Computer Science. We are particularly interested in identifying the sections in the paper where these citations appear, as we hypothesize that this might signal different types of interdisciplinary engagement.

To do this, we extract the content of the each article using Grobid[†]. Then, we use the SciPDF Parser[‡] to convert the extracted content into a dictionary format including the article text and citations, as well as section breakdowns. We model each publication as a set of canonical section headers typical of NLP papers paired with their corresponding in-text citations. We consider the following canonical sections: “Introduction”, “Related Work”, “Method”, “Experiments”, “Conclusion” and “Appendix”. To arrive at this canonical section breakdown, the sections initially parsed by Grobid are assigned to one of these headers using a rule-based approach based on string matching. Then, all in-text citations within a section are mapped to the corresponding “canonical” section. Details about this process are outlined in Appendix A.2.

In addition, information for each entry in the “References” section of a publication (title, journal, publication data, id, etc.) is extracted. For all references with an available, valid id (DOI, arXiv, or url), we query the Semantic Scholar API (Kinney et al., 2023) for a “field of study”. To handle entries with no valid id, we match venues to the appropriate field of study using Google Scholar’s lists of venues per field. All remaining unassigned venues were annotated manually by the authors of the paper. Finally, all in-text citations were mapped to reference entries by string matching the author name and the publication year.

3.4 Argument Analysis

Once we have identified all interdisciplinary references, we are interested in modeling the argumentative role they play in the article. To do this, we build on the argumentation schema proposed by Lauscher et al. (2018). This schema builds on the

[†]<https://github.com/kermitt2/grobid>

[‡]https://github.com/titipata/sci_pdf_parser

Toulmin model (Toulmin, 2003), one of the most widely used theoretical frameworks of argumentation. The Toulmin model was originally conceived for the legal domain, and recognizes six types of argumentative components: claim, data, warrant, backing, qualifier, and rebuttal. Lauscher et al. (2018) do an annotation study on scientific documents, and simplify the Toulmin model by focusing only on claims and data. Further, they break down claims into *own claims* and *background claims* to differentiate between claims that relate to the author’s own work, and claims that relate to the background of the author’s work (e.g., about related work or common practices).

We build a classifier to identify whether interdisciplinary references correspond to *data* (e.g., fact or example that serve as evidence for or against a claim), *background claims*, or *own claims*. To do so, we fine-tune a pre-trained BERT classifier using the dataset provided by Lauscher et al. (2018). This dataset is comprised of 13,592 arguments: 6,004 own claims 3,291 background claims, and 4,297 data statements. Upon closer inspection of the dataset, we found that many examples correspond to figure titles and standalone citations. To deal with this, we removed all training examples with no verb phrases using spaCy. This results in 8,737 arguments: 4,968 own claims, 2,613 background claims, and 1,156 data statements.

4 Experiments and Analysis

To validate our framework, we first perform an experimental evaluation of each component. Then, we use the full framework to perform an extensive analysis of interdisciplinary engagement for our full dataset of CSS+NLP papers.

4.1 Framework Evaluation

To train and evaluate the track identification and argument type classifiers, we performed 5-fold cross-validation and trained using the AdamW optimizer, the cross-entropy loss, and a learning rate of $1e - 5$. For early stopping, we used the macro F1 on the validation set, consisting of 20% of the training examples for each fold.

We present the resulting metrics of our track classifier in Tab. 1. We obtain strong performance for this task considering the highly imbalanced nature of the data. This suggests that we can trust that our dataset of NLP+CSS papers is representative

Class	Precision	Recall	F1
CSS	0.8189	0.8326	0.8257
Not CSS	0.9742	0.9716	0.9729
Macro Avg	0.8965	0.9021	0.8993

Table 1: Avg. Results for Track Prediction

Class	Precision	Recall	F1
Own Claim	0.7460	0.7693	0.7564
Background Claim	0.6289	0.6184	0.6218
Data	0.5233	0.4669	0.4745
Macro Avg	0.6327	0.6182	0.6176

Table 2: Avg. Results for Argument Type Prediction

of the real distribution. Detailed results per fold are shown in Tab. 8 in Appendix A.3.

We present the resulting metrics of our argument type classifier in Tab. 2. We obtain relatively good performance for the two types of claims, but struggle with data statements. This is a challenging task, and our results are in line with the skewed nature of the dataset, where there is considerably less support for data examples. Detailed results per fold are shown in Tab. 10 in Appendix A.3.

4.2 Analysis of Interdisciplinary Engagement

In this section, we use the framework introduced in Sec. 3 to perform an exhaustive analysis of the engagement of NLP+CSS papers with work outside of NLP, Computer Science and Linguistics. To do this, we first used the CSS track classifier to derive the dataset presented in Fig 1. Then, we ran the citation mapping procedure. Next, for every sentence involving or preceding a citation or reference, we predict its argument type using our argument type classifier. To train the final argument type classifier, we used the full dataset of arguments from Lauscher et al. (2018), containing all five folds. Finally, we model 15 topics in the abstracts of the NLP+CSS papers to identify growing and shrinking trends.

The final dataset is comprised of 741 NLP+CSS papers, published across five NLP conferences and one workshop, and spanning 9 years (2014-2023). Within these 741 publications, we have a total of 16,652 references annotated with the canonical section in which they appear, their scientific discipline and their predicted argument type.

Below, we present our analysis organized by the

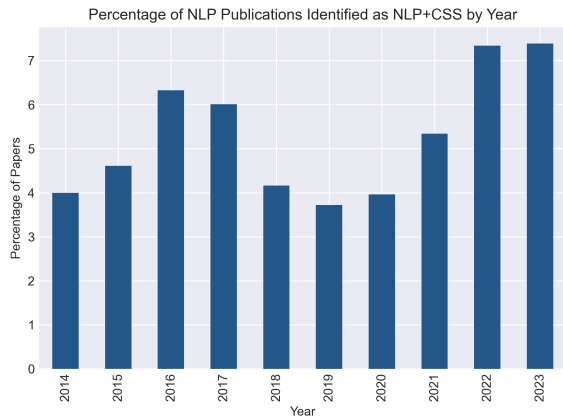


Figure 2: Percentage of all gathered NLP papers that were predicted or labelled as NLP+CSS papers

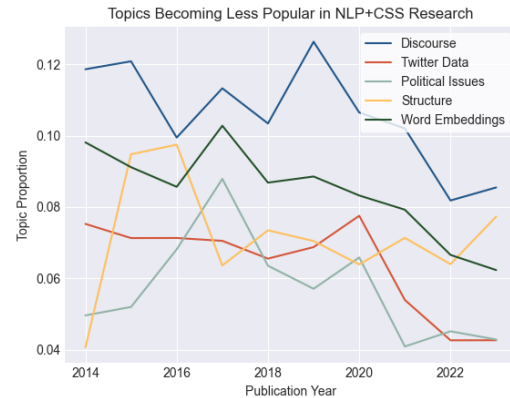


Figure 3: Topics becoming less popular in NLP+CSS research

main trends that we observed in the data.

NLP+CSS research is growing. Fig. 2 shows the number of NLP+CSS papers per year in all five NLP conferences. In this figure, we observe that the percentage of NLP work identifying as NLP+CSS research has significantly increased in the last three years. This growth comes at a time where the volume of NLP work in general has significantly increased. We also observe a peak in NLP+CSS work in 2016 and 2017.

Rising and falling topic trends might explain NLP+CSS prominence. To gain a better understanding of shifts in content of the NLP+CSS papers, we conducted topic modelling with gibbs sampling over all abstracts using tomotopy (Lee, 2022). We modelled 15 topics and eliminated the 50 most common words. Based on the top 10 words associated with each topic, as well as the top documents associated with each topic, we manually assigned the topic a title. Details for each of the 15 topics, including associated words and assigned title are presented in App. A.5.

We used topic distribution information to understand shifts in topics over the years. We identify five topics that have become less prevalent over the last five years (Fig. 3), and three topics that have become increasingly popular over the full 9 year period (Fig. 4). First, we can appreciate that the 2016–2017 peak in NLP+CSS papers corresponds with increased interest in topics related to politics, public discourse and hate speech. We hypothesize that these trends could be related to the U.S. general election and the uptake in political discourse on social media. A similar peak can be seen for political issues around 2020, when the next U.S.

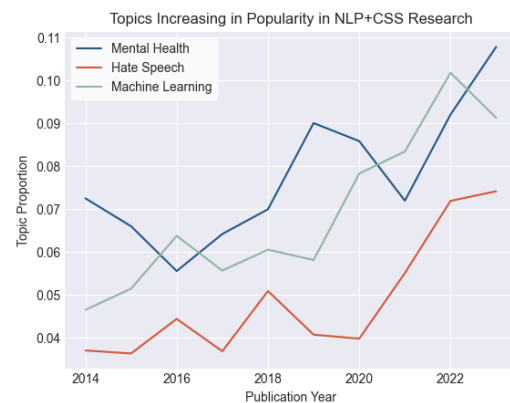


Figure 4: Topics becoming more popular in NLP+CSS research

general election occurred.

We also observe a decline in techniques like structured prediction after 2016, when neural networks like LSTMs became more popular. Word embeddings experienced a pronounced peak between 2018 and 2020—coinciding with the rise of contextualized word representations—and it has been decreasing since then. We hypothesize that this is influenced by a shift to LLMs and generative solutions. On the other hand, general machine learning vocabulary has been steadily increasing in the last 9 years. This is unsurprising, as NLP research has become increasingly more entangled with machine learning research. However, this has not caused a shift in focus away from social topics, as we also observe a steady incline in research around mental health and hate speech.

Finally, we observe a sharp decrease in papers dealing with Twitter data in the last two years. This coincides with changes in leadership at Twitter, and

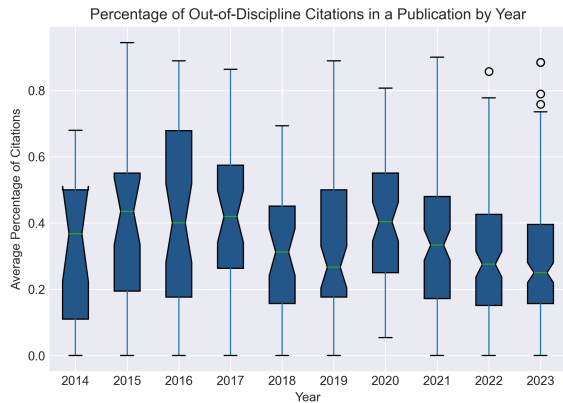


Figure 5: Average % of Out-of-Discipline citations in NLP+CSS papers per year

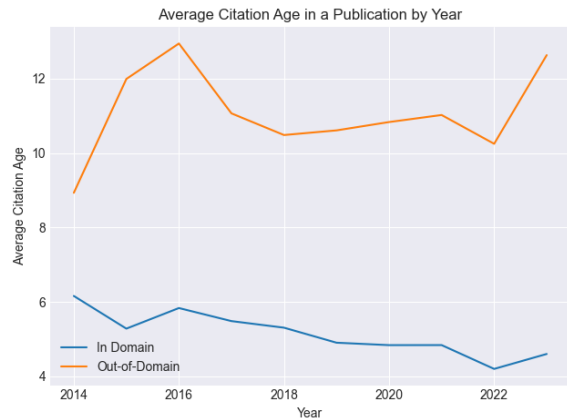


Figure 6: The average age of Out-of-Discipline citations are significantly older at the time of reference

the increasing difficulty in accessing the data.

Cross-disciplinary engagement in NLP+CSS articles has decreased in the last three years. While the number of NLP+CSS papers has grown in the last five years, the proportion of out-of-discipline and specifically social scientific papers cited has decreased in the same time span (Figs 5 and 7). Initially, peaks in the amount of cross-disciplinary engagement coincide with peaks in the prevalence of NLP+CSS work (around 2016 and 2020, as seen on Fig. 2). However, after 2020, the trends seem to be in opposition to each other - more NLP+CSS work but less interdisciplinary engagement. Interestingly, the decrease in proportion of interdisciplinary citations coincides with the LLM boom in NLP research.

To further inspect this trend, we look at the most cited papers in NLP+CSS work for each year (Tab 3). The most cited papers prior to 2018 had to do with social-adjacent topics such as dialog (Prabhakaran and Rambow, 2014) and language in social media data (Mitchell et al., 2015), and were not as frequently cited (cited in 11-15 in-text citations). Starting in 2018, top-cited papers include a survey of affective computing (Poria et al., 2017), a paper about deep neural networks (Alzantot et al., 2018), and pre-trained language models (Liu et al., 2019). The amount of papers citing them significantly grew (20-90 in-text citations).

Out-of-Discipline citations are older. In plotting the average age of Out-of-Discipline versus In-discipline citations at the age of reference (Fig. 6), we find that Out-of-Discipline citations are significantly older. This may communicate a tendency to engage only with more well-known, seminal

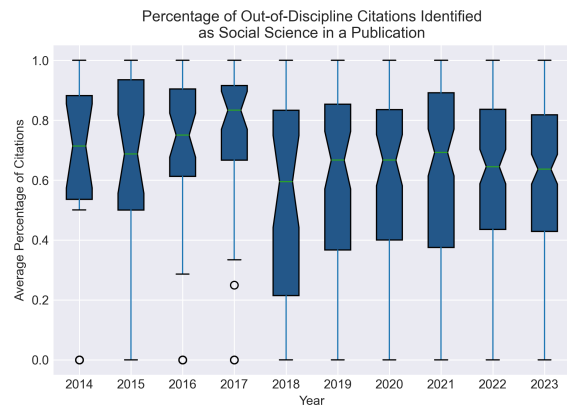


Figure 7: Average percentage of Out-of-Discipline citations within a publication that are considered Social Science-specific

papers. It is also illustrative of the much shorter “life-span” of AI-adjacent research. Over the past decade, this has become increasingly true (Singh et al., 2023; Nguyen and Eger, 2024).

Dedicated workshops are more interdisciplinary. In Fig. 8 we show boxplots for the proportion of interdisciplinary citations per venue. We observe no meaningful difference in the interdisciplinary citation patterns of CSS papers between the different NLP conferences, with the exception of AACL, which appears to be less interdisciplinary. This could be due to the fact that AACL is a new venue, and therefore attracts more traditional NLP work.

However, we find a significant increase in interdisciplinary citations for the NLP+CSS workshop. This suggests that dedicated workshops attract more interdisciplinary work, which is in line with the mission of the workshop series: to foster the progress of CSS, and to integrate CSS with

Year	Most Cited Paper	Times Cited
2014	Predicting Power Relations between Participants in Written Dialog from a Single Thread (Prabhakaran and Rambow, 2014)	11
2015	Exploiting Similarities among Languages for Machine Translation (Mikolov et al., 2013)	10
2016	Inferring Latent User Properties from Texts Published in Social Media (Volkova et al., 2015)	15
2017	Quantifying the Language of Schizophrenia in Social Media (Mitchell et al., 2015)	12
2018	A review of affective computing: From unimodal analysis to multimodal fusion (Poria et al., 2017)	16
2019	Generating Natural Language Adversarial Examples (Alzantot et al., 2018)	19
2020	RoBERTa: A Robustly Optimized BERT Pretraining Approach (Liu et al., 2019)	20
2021	RoBERTa: A Robustly Optimized BERT Pretraining Approach (Liu et al., 2019)	39
2022	RoBERTa: A Robustly Optimized BERT Pretraining Approach (Liu et al., 2019)	61
2023	RoBERTa: A Robustly Optimized BERT Pretraining Approach (Liu et al., 2019)	90

Table 3: Most cited paper per year

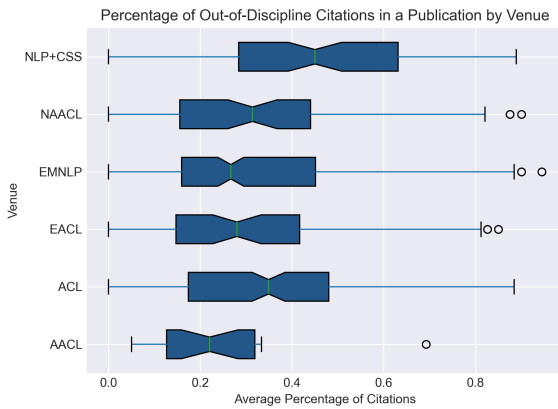


Figure 8: % of Out-of-Discipline citations per venue

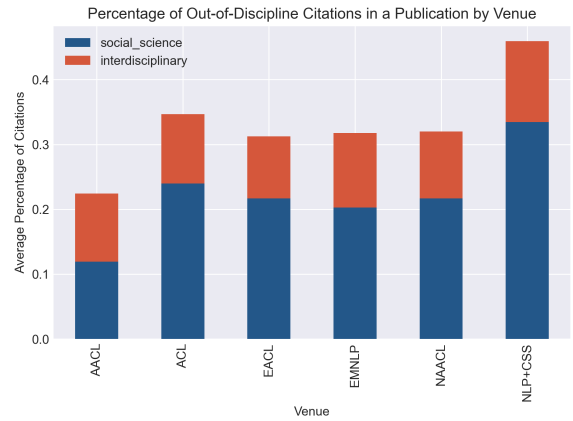


Figure 9: % Out-of-Discipline per venue

current trends and techniques in NLP.

Social science papers are the most commonly cited. Unsurprisingly, among the number of out-of-discipline papers cited in CSS papers, the vast majority correspond to social science venues. This can be observed in Fig. 7 (prop. of social science papers among out-of-discipline papers per year) and Fig. 9 (prop. of social science papers among out-of-discipline papers per venue). Moreover, we find that the trend is very stable among different years and venues. To perform this analysis, we consider the social science and humanities fields outlined in App A.4.

Finally, to further investigate which social science (and humanities) fields are most commonly cited, we plot the top-10 fields in Fig 10. We find that *psychology* is the top most cited social science / humanities field. It is followed by *political science*, *general social science*, *sociology*, *business*, *economics*, *communication*, and *education*.

In-discipline references are used more often for own claims and data statements. In Tab. 4 we can appreciate the differences in argument type between in-discipline and out-of-discipline refer-

Citation Type	Background Claim	Data	Own Claim
In-Discipline	53.98%	6.92%	39.1%
Out-of-Discipline	64.16%	4.77%	31.06%
Social Science	68.91%	4.13%	26.96%

Table 4: Percentage of Argument Types Supported by In-Discipline, Out-of-Discipline and Social Science-specific Out-of-Discipline Citations

ences. We find that when making claims that relate to the author’s own work (own claims), and stating facts or examples that serve as evidence, authors more often reference work within the same discipline. Conversely, when making claims that relate to the background of their work (background claims), authors more often reference work outside of their discipline. We also note that this difference is even more pronounced in Social Science-specific Out-of-Discipline citations. While this is an unsurprising result, it is interesting that there is still a significant amount of out-of-discipline citations used to make *own claims*, which signals meaningful interdisciplinary engagement.

Sections matter when referencing out-of-discipline work. In Fig. 11 we can appreciate significant differences in argument types and citation

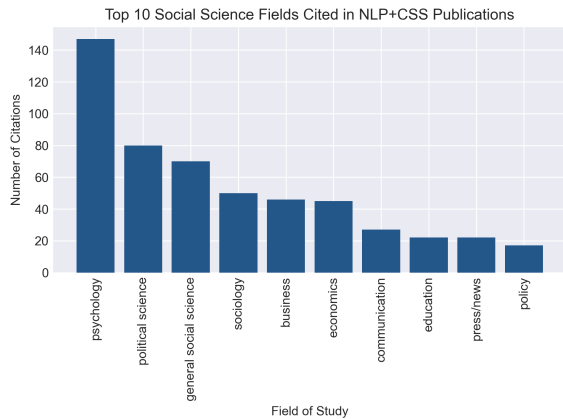


Figure 10: Top 10 most cited social science fields

frequency by section. The number of background claims made in the introduction increases significantly for out-of-discipline references. This makes sense for CSS papers that are attempting to frame and motivate their contributions with respect to the broader social science literature. Additionally, we can see that for both method and experiments sections, in-discipline references are used more often to back *own claims*, while out-of-discipline references are used more evenly to support both *own claims* and *background claims*.

Papers with higher rates of Out-of-Discipline citations are integrative. We conduct a qualitative analysis of the three papers identified as having the greatest proportion of Out-of-Domain references and compare to the three papers with the lowest proportion. Our observations for each paper are summarized in Tab. 5.

We find that the papers identified with the most Out-of-Discipline citations each seek to build upon existing social science work, tying relevant theory strongly into their motivation, methods, and discussion of results. Following (McCarthy and Dore, 2023), we are inclined to describe them as “integrative” papers. Alternatively, we find that papers with the lowest rates of Out-of-Discipline citations are mainly method papers grounded in computer science research. Each of these papers addresses a relevant social issue or task, but the main focus is on formulating a prediction task, proposing a computational model and analyzing the prediction performance.

5 Discussion and Future Work

We emphasize that this paper represents only a first step toward our envisioned framework. Go-

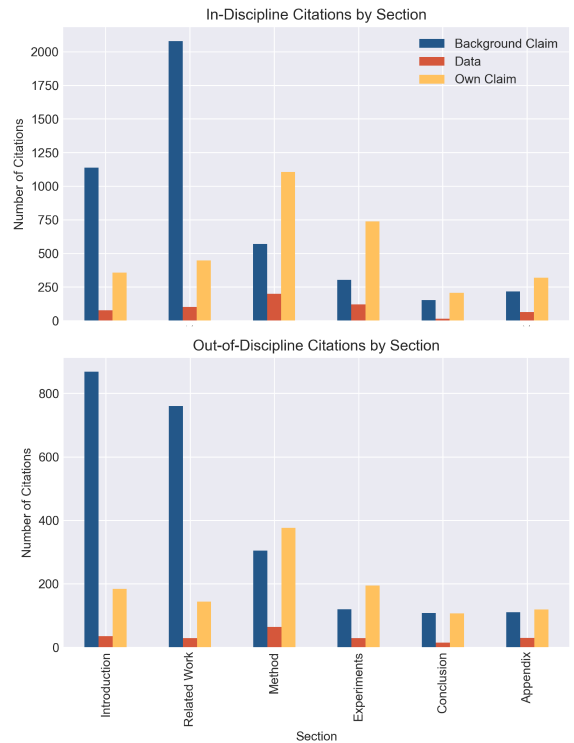


Figure 11: Argument Types Supported by In-discipline vs. Out-of-Discipline citations per Section

ing forward, we will expand our labelled dataset for greater coverage of additional tracks and workshops typical of NLP conferences. We aim to predict, with high confidence, what track a paper was submitted to, allowing for a more holistic analysis of the field of NLP and its relationship with out-of-discipline work. We will conduct both general and per-track analyses to identify additional trends and explore whether the trends observed in NLP+CSS hold true in other types of work.

We are particularly interested in studying the lasting power of NLP work and how that has shifted under recent advancements in the field. We are also interested in expanding on (McCarthy and Dore, 2023) to further investigate the differences in integrative and descriptive work and their engagement patterns with out-of-discipline scholarship in the field of NLP as a whole. We seek to provide examples of how integrative work may be carried out in the age of LLMs.

In addition to expanding beyond NLP+CSS, we are also interested in improving on our argument analysis component. We want to experiment with alternative frameworks, as well as with modeling argumentative relations between claims. Moreover, we will further investigate our hypothesis that the

Most Interdisciplinary	
(Wu et al., 2023)	This paper presents a study of values, morals and gender bias in a dataset of folk tales spanning cultures. Most experiments are designed to reinforce a relevant theoretical claim.
(Yang et al., 2015)	This is a study of the relationship between personal traits and brand preference. The method is grounded in theory and seeks to investigate a debated topic in marketing research through a large scale study.
(CH-Wang and Jurgens, 2021)	This paper studies differences in word choices for significant others and indefinite people to examine shifts in societal attitudes toward sexuality and gender. Motivation for their method and discussion of results relies heavily on theoretical background.
Least Interdisciplinary	
(Yang et al., 2022)	This paper presents a method for using facial emotions to extract sentiment from sentence-image pairs.
(Zhou et al., 2016)	This paper studies enhancing limited personal usage information with the goal of improved search personalization. They mainly call on previous computational methods to motivate design choices.
(Li et al., 2023)	This paper examines the effectiveness of identifying stance in social media posts given background knowledge about the topic. Their proposed method is a variant of Retrieval Augmented ChatGPT.

Table 5: Summary of observations from analysis of NLP+CSS publications with greatest and least proportion of Out-of-Discipline citations

location of a citation within a paper may signal higher or lower engagement. Conducting qualitative analysis with a science-of-science expert will allow us to tie these observations to meaningful differences in engagement. We envision combining this information with the argumentation framework to make deeper claims about the way citations are used within a publication. We also recognize the need to integrate and contrast our framework with recent Science of Science techniques such as citation network analysis.

6 Conclusion

We recognize a need to evaluate interdisciplinary research due to its growing popularity paired with a lack of sufficient methods for studying engagement between disciplines. In line with this, we propose a general scalable framework for tracking interdisciplinary citations within publications. Our framework allows for tracking the section where interdisciplinary citations appear and the argumentative role they play within a publication.

To showcase our framework, we performed a large scale analysis on the interdisciplinary engagement of research in the field of natural language processing and computational social science. To do this, we constructed a comprehensive dataset of NLP+CSS papers published in the NLP venues in the last decade. Our analysis revealed several trends, including a growth in the number of NLP+CSS publications, compounded with a decrease in cross-disciplinary engagement in NLP+CSS papers coinciding with the advent of LLMs. These findings are in line with previous work highlighting the gaps between the two

fields (Baden et al., 2022; McCarthy and Dore, 2023). However, we also find that dedicated workshops, such as the *NLP and Computational Social Science workshop*, attract contributions that exhibit higher engagement with the social scientific literature.

7 Limitations

The work presented in this paper has four main limitations: (1) We defined interdisciplinary references as those that cited a paper outside of NLP, Computer Science or Linguistics. We recognize that this is a simplification and that scientific contributions can vary widely within certain venues. (2) The classifiers used to identify argument types was trained and evaluated on out-of-domain data. While this data was also comprised of scientific articles, some domain drift is to be expected when moving from the computer graphics domain to the natural language processing domain. A post-hoc manual evaluation is needed to check and establish the performance for our dataset. (3) We complemented our citation analysis with a topic analysis to tie the findings to some of the most prominent research trends in the literature. We recognize the limitation of topic models to accurately capture this type of information accurately. However, we believe that this risk is diminished when looking at aggregated trends, rather than at individual mapping between papers and topics. (4) The fact that we are using automated techniques for the analysis necessarily carries some uncertainty. Even if we were to improve our models considerably, our large-scale analysis has a margin of error. It is important to acknowledge this when presenting our

findings.

8 Ethical Considerations

To the best of our knowledge, no code of ethics was violated during the development of this project. We used publicly available tools and data to develop our framework and perform our analysis. We reported all pre-processing steps, learning configurations, hyperparameters, and additional technical details. Due to space constraints, some information was relegated to the Appendix. The results reported in this paper support our claims and we believe that they are reproducible. The analysis reported in Section 4.2 was done using the outputs of matching algorithms and machine learning techniques and do not represent the authors personal views. The uncertainty of our predictions was adequately acknowledged in the Limitations Section, and the estimated accuracy was reported in Sec. 4.1.

References

- Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. [Generating natural language adversarial examples](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.
- Awais Athar and Simone Teufel. 2012. Context-enhanced citation sentiment detection. In *Proceedings of the 2012 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*, pages 597–601.
- Christian Baden, Christian Pipal, Martijn Schoonvelde, and Mariken A. C. G van der Velden. 2022. [Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda](#). *Communication Methods and Measures*, 16(1):1–18.
- Arne Binder, Leonhard Hennig, and Bhuvanesh Verma. 2022. [Full-text argumentation mining on scientific publications](#). In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*, pages 54–66, Online. Association for Computational Linguistics.
- Sky CH-Wang and David Jurgens. 2021. [Using sociolinguistic variables to reveal changing attitudes towards sexuality and gender](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9918–9938, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Shiji Chen, Junping Qiu, Clément Arsenault, and Vincent Larivière. 2021. [Exploring the interdisciplinary patterns of highly cited papers](#). *Journal of Informetrics*, 15(1):101124.
- Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. [Specter: Document-level representation learning using citation-informed transformers](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282.
- Yingqiang Gao, Nianlong Gu, Jessica Lam, and Richard H.R. Hahnloser. 2022. [Do discourse indicators reflect the main arguments in scientific papers?](#) In *Proceedings of the 9th Workshop on Argument Mining*, pages 34–50, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.
- Aaron Gerow, Yuening Hu, Jordan Boyd-Graber, David M Blei, and James A Evans. 2018. [Measuring discursive influence across scholarship](#). *Proceedings of the national academy of sciences*, 115(13):3308–3313.
- Justin Grimmer and Brandon M. Stewart. 2013. [Text as data: The promise and pitfalls of automatic content analysis methods for political texts](#). *Political Analysis*, 21(3):267–297.
- Edward J. Hackett. 2000. [12. interdisciplinary research initiatives at the u.s. national science foundation](#). In Peter Weingart and Nico Stehr, editors, *Practising Interdisciplinarity*, pages 248–259. University of Toronto Press.
- David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. [Measuring the evolution of a scientific field through citation frames](#). *Transactions of the Association for Computational Linguistics*, 6:391–406.
- Rodney Michael Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David W. Graham, F.Q. Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler Murray, Christopher Newell, Smita R Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, A. Tanaka, Alex D Wade, Linda M. Wagner, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine van Zuylen, and Daniel S. Weld. 2023. [The semantic scholar open data platform](#). *ArXiv*, abs/2301.10140.
- Suchetha Nambanoor Kunnath, David Pride, and Petr Knoth. 2022. [Dynamic context extraction for citation classification](#). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International*

- Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 539–549.
- Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. 2018. [An argument-annotated corpus of scientific publications](#). In *Proceedings of the 5th Workshop on Argument Mining*, pages 40–46, Brussels, Belgium. Association for Computational Linguistics.
- John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.
- Erin Leahey and Sondra N. Barringer. 2020. [Universities’ commitment to interdisciplinary research: To what end?](#) *Research Policy*, 49(2):103910.
- Minchul Lee. 2022. [bab2min/tomotopy: 0.12.3](#).
- Ang Li, Bin Liang, Jingqian Zhao, Bowen Zhang, Min Yang, and Ruifeng Xu. 2023. [Stance detection on social media with background knowledge](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15703–15717, Singapore. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized BERT pretraining approach](#). *CoRR*, abs/1907.11692.
- Li Lucy, Jesse Dodge, David Bamman, and Katherine Keith. 2023. [Words as gatekeepers: Measuring discipline-specific terms and meanings in scholarly publications](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6929–6947, Toronto, Canada. Association for Computational Linguistics.
- Arya D. McCarthy and Giovanna Maria Dora Dore. 2023. [Theory-grounded computational text analysis](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1586–1594, Toronto, Canada. Association for Computational Linguistics.
- Kathy McKeown, Hal Daume III, Snigdha Chaturvedi, John Paparrizos, Kapil Thadani, Pablo Barrio, Or Biran, Suvarna Bothe, Michael Collins, Kenneth R Fleischmann, et al. 2016. Predicting the impact of scientific concepts using full-text features. *Journal of the Association for Information Science and Technology*, 67(11):2684–2696.
- Óscar E Mendoza, Wojciech Kusa, Alaa El-Ebshihy, Ronin Wu, David Pride, Petr Knuth, Drahomira Herrmannova, Florina Piroi, Gabriella Pasi, and Allan Hanbury. 2022. Benchmark for research theme classification of scholarly documents. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 253–262.
- Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. [Exploiting similarities among languages for machine translation](#). *ArXiv*, abs/1309.4168.
- Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. [Quantifying the language of schizophrenia in social media](#). In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 11–20, Denver, Colorado. Association for Computational Linguistics.
- Saif M. Mohammad. 2020. [Examining citations of natural language processing literature](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5199–5209, Online. Association for Computational Linguistics.
- Tsendsuren Munkhdalai, John P Lalor, and Hong Yu. 2016. Citation analysis with neural attention models. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*, pages 69–77.
- Suchetha N. Kunnath, David Pride, Drahomira Herrmannova, and Petr Knuth. 2021. [Overview of the 2021 SDP 3C citation context classification shared task](#). In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 150–158, Online. Association for Computational Linguistics.
- Hoa Nguyen and Steffen Eger. 2024. [Is there really a citation age bias in nlp?](#) *ArXiv*, abs/2401.03545.
- Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.
- Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. [A review of affective computing: From unimodal analysis to multimodal fusion](#). *Information Fusion*, 37:98–125.
- Alan L. Porter and Ismael Rafols. 2009. [Is science becoming more interdisciplinary? Measuring and mapping six research fields over time](#). *Scientometrics*, 81(3):719–745.
- Vinodkumar Prabhakaran, William L Hamilton, Dan McFarland, and Dan Jurafsky. 2016. Predicting the rise and fall of scientific topics from trends in their rhetorical framing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1170–1180.
- Vinodkumar Prabhakaran and Owen Rambow. 2014. [Predicting power relations between participants in written dialog from a single thread](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 339–344, Baltimore, Maryland. Association for Computational Linguistics.

Vahed Qazvinian and Dragomir Radev. 2008. Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 689–696.

Diana Rhoten and Andrew Parker. 2024. [Risks and Rewards of an Interdisciplinary Research Path](#). *Science*, 306(5704):2046.

Mukund Rungta, Janvijay Singh, Saif M. Mohammad, and Diyi Yang. 2022. [Geographic citation gaps in NLP research](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1371–1383, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Janvijay Singh, Mukund Rungta, Diyi Yang, and Saif Mohammad. 2023. [Forgotten knowledge: Examining the citational amnesia in NLP](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6192–6208, Toronto, Canada. Association for Computational Linguistics.

Sonita Te, Amira Barhoumi, Martin Lentschat, Frédérique Bordignon, Cyril Labbé, and François Portet. 2022. Citation context classification: Critical vs non-critical. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 49–53.

Stephen E. Toulmin. 2003. *The Uses of Argument*. Cambridge University Press.

V. Kelly Turner, Karina Benessaiah, Scott Warren, and David Iwaniec. 2015. [Essential tensions in interdisciplinary scholarship: navigating challenges in affect, epistemologies, and structure in environment–society research centers](#). *Higher Education*, 70(4):649–665.

Richard Van Noorden. 2015. [Interdisciplinary research by the numbers](#). *Nature*, 525:306–7.

Svitlana Volkova, Yoram Bachrach, Michael Armstrong, and Vijay Sharma. 2015. [Inferring latent user properties from texts published in social media](#). In *AAAI Conference on Artificial Intelligence*.

Wilhelm Vosskamp, Raymond C Miller, and Julie Thompson Klein. 1986. From scientific specialization to the dialogue between the disciplines. *Issues in Interdisciplinary Studies*.

Winston Wu, Lu Wang, and Rada Mihalcea. 2023. [Cross-cultural analysis of human values, morals, and biases in folk tales](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5113–5125, Singapore. Association for Computational Linguistics.

Chao Yang, Shimei Pan, Jalal Mahmud, Huahai Yang, and Padmini Srinivasan. 2015. [Using personal traits for brand preference prediction](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 86–96, Lisbon, Portugal. Association for Computational Linguistics.

Hao Yang, Yanyan Zhao, and Bing Qin. 2022. [Face-sensitive image-to-emotional-text cross-modal translation for multimodal aspect-based sentiment analysis](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3324–3335, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dani Yogatama, Michael Heilman, Brendan O’Connor, Chris Dyer, Bryan R Routledge, and Noah A Smith. 2011. Predicting a scientific community’s response to an article. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 594–604.

Dong Zhou, Séamus Lawless, Xuan Wu, Wenyu Zhao, and Jianxun Liu. 2016. [Enhanced personalized search using social data](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 700–710, Austin, Texas. Association for Computational Linguistics.

A Appendix

A.1 Detailed Dataset Statistics

In Figure 12 we present the full dataset, including labelled and unlabelled data, by publication year. Figure 13 shows this dataset broken down by the conference or workshop in which it was published.

Figure 14 shows the labelled subset of data broken down by year and label. Figure 15 breaks this subset down by publication conference/workshop.

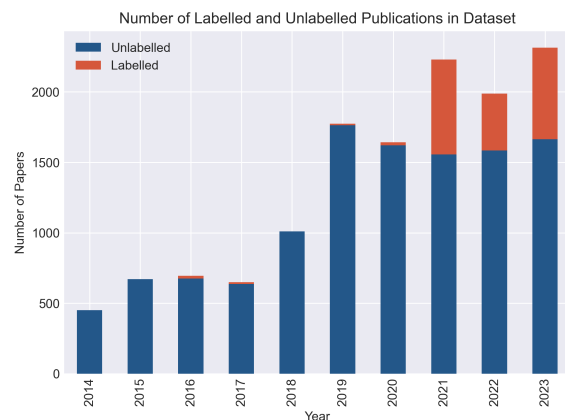


Figure 12: Full unlabelled dataset by year

A.2 Details on Citation Mapping

The Grobid and SciPDF pipeline result in a dictionary format including the article text and citations, as well as section breakdowns. To assign canonical section titles to each section, we string matched on a set of possible titles common in publications associated with our predetermined section titles. These are included in Tab. 6.

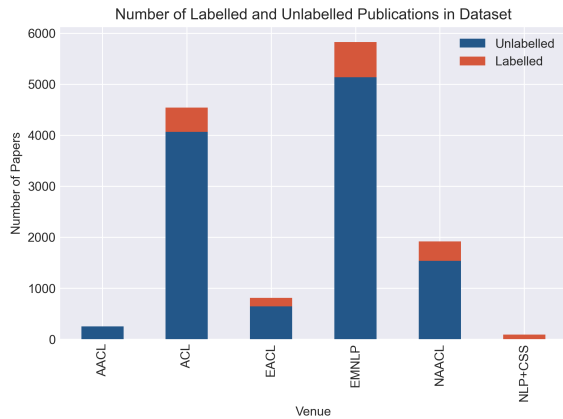


Figure 13: Full unlabelled dataset by conference

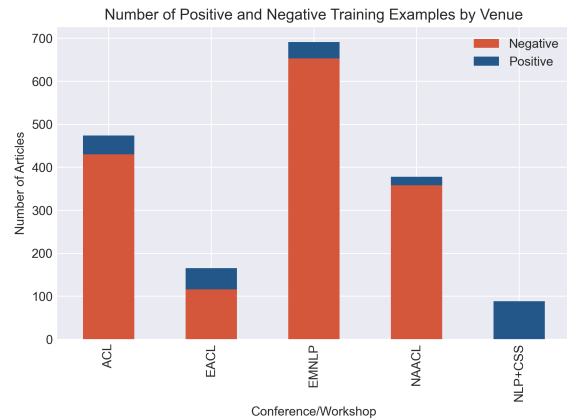


Figure 15: Labeled dataset by conference

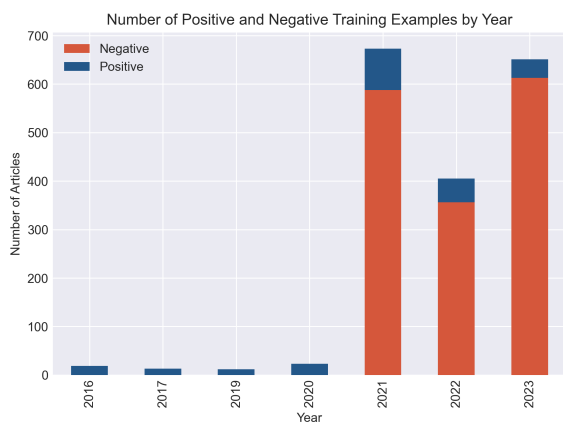


Figure 14: Labeled dataset by year

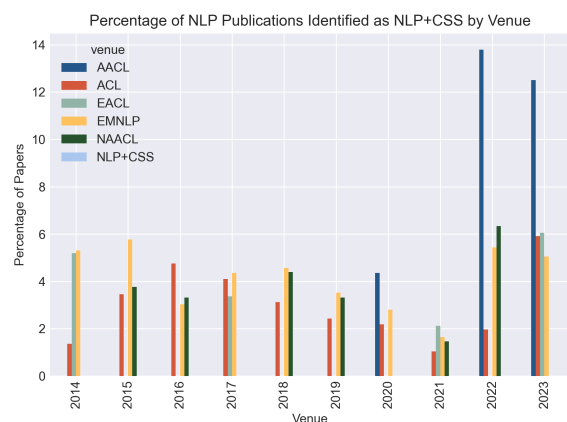


Figure 16: Percentage of CSS papers by year and venue

The dictionary includes in-text citations for every section. We matched these to an entry in the references section by string matching on first author last name and publication date.

A.3 Classifier Data Details

In Tab. 7 we show the number of examples present in each split during 5-fold cross-validation for the track classifier.

In Tab. 9 we show the number of examples present in each split during 5-fold cross-validation for the argument classifier.

A.4 Social Science Fields

accounting, anthropology, applied psychology, art, arts and humanities (miscellaneous), biological psychiatry, business, communication, criminology, cultural studies, demography, developmental and educational psychology, ecology, evolution, behavior and systematics, economics, economics and econometrics, economics, econometrics and

finance (miscellaneous), education, environmental science, epidemiology, ethics, experimental and cognitive psychology, finance, gender studies, general health science, general humanities, general psychology, general social science, genetics, geography, geography, planning and development, health informatics, health information management, health policy, history, history and philosophy of science, humanities, informatics, information science, information systems, information systems and management, language and linguistics, library and information science, life-span and life-course studies, linguistics, linguistics and language, literature, marketing, pediatrics, perinatology and child health, philosophy, policy, political science, political science and international relations, press/news, psychiatry and mental health, psychology, public administration, public health, rehabilitation, social psychology, social science (miscellaneous), sociology, sociology and political science, speech and hearing, strategy and management

Canonical Title	Matching Sections
Introduction	introduction
Related Work	related work, background, related research
Method	method, methodology, approach, notation, technique, algorithm, architecture, design, solution, method, approach, methodology, model, corpus, data
Experiments	experiment, evaluation, test, analysis, compare, accuracy, scores, our method vs., state-of-the-art, baseline, results, performance, evaluation
Conclusion	conclusion, future work, discussion, limitation, ethical consideration, ethics
Appendix	appendix

Table 6: Paper section titles mapped to our canonical titles

Fold	Train	Validation	Test
0	1145	288	358
1	1145	288	358
2	1145	288	358
3	1147	286	358
4	1145	287	359

Table 7: Number of examples in each data split used for 5-fold cross-validation with task classifier

Fold	CSS	Not CSS	Macro
0	0.8872	0.9354	0.9093
1	0.9482	0.9482	0.9482
2	0.8799	0.8908	0.8852
3	0.8701	0.8916	0.8804
4	0.9031	0.8589	0.8790
Avg	0.9136	0.9362	0.9249
Stdev	0.03194	0.0158	0.0238

Table 8: F1 for CSS Track Prediction per Fold

A.5 Topic Modelling Details

We conducted topic modelling with tomotopy (Lee, 2022) over all abstracts. We modelled 15 topics and eliminated the 50 most-common words. Based on the top 10 words associated with each topic, we manually assigned the topic a title. The top ten words associated with each topic and the manually assigned label are included in Tab. 11.

Fold	Train	Validation	Test
0	5313	1537	1887
1	5608	1567	1562
2	5657	1498	1582
3	5303	1779	1655
4	5478	1208	2051

Table 9: Number of examples in each data split used for 5-fold cross-validation with argument classifier

Fold	Own Claim	Background Claim	Data	Macro
0	0.7989	0.6508	0.5	0.6499
1	0.7031	0.6251	0.4367	0.5883
2	0.7652	0.5957	0.4389	0.5999
3	0.763	0.6405	0.5059	0.6365
4	0.7516	0.597	0.491	0.6132
Avg	0.7564	0.6218	0.4745	0.6176
Stdev	0.0346	0.0250	0.0339	0.0255

Table 10: F1 for Argument Type Prediction per Fold

Topic Name	Top 10 Words
Discourse	context, only, but, at, time, also, discourse, study, online, evidence
Sarcasm	al, sarcasm, et, moral, its, aspect, have, methods, been, then
State of the Art	over, approach, stateoftheart, baseline, present, outperforms, stance, novel, based, annotated
Emotion Detection	emotion, knowledge, information, datasets, demonstrate, experiments, multimodal, stateoftheart, effectiveness, rumor
Twitter Data	users, user, twitter, tweets, posts, predict, emotional, methods, individuals, studies
Mental Health	learning, framework, existing, health, mental, large, tasks, natural, novel, proposed
Gender	more, study, gender, than, nlp, have, may, find, groups, people
Political Issues	political, how, identify, us, computational, articles, through, identifying, science, issues
Hate Speech	speech, content, hate, not, online, detecting, one, also, but, personal
Semantic Structure	semantic, used, information, annotated, human, structure, them, documents, sentences, set
Linguistics	features, new, research, languages, linguistic, into, english, how, across, while
Word Embeddings	approach, words, word, embeddings, use, method, all, same, two, predicting
Events	eg, information, about, prediction, both, events, methods, also, change, event
Machine Learning	training, bias, classification, datasets, trained, both, at, through, been, problem
Conversations	conversations, conversation, strategies, computational, power, persuasion, where, framework, not, conversational

Table 11: 15 topics identified in abstracts with hand-labelled titles

Author Index

- Abadi, David, 1
Abdurahaman, Zara Fatima, 68
Acharya, Anurag, 46
Acuna, Daniel, 144
Alcock, Keith, 68
Alexeeva, Maria, 68
Arora, Abhishek, 99
- Baldwin, Timothy, 16
Broek, Tijs A. Van Den, 1
Bryan, Tom, 99
- Curto, Georgina, 32
- Dahal, Shreeja, 46
Daoud, Adel, 57
Dell, Melissa, 99
Dumitru, Razvan-Gabriel, 68
- Estrada, Diego, 46
- Finlayson, Mark, 46
Franklin, Brevin, 99
Fraser, Kathleen C., 32
Frermann, Lea, 16
- Gomez, Diana, 46
- Hoyle, Alexander, 144
- Jeong, Cheonkam, 68
Jerzak, Connor Thomas, 57
Johansson, Moa, 57
Johansson, Richard, 57
- Kirchhoff, Brian, 68
Kiritchenko, Svetlana, 32
- Leto, Alexandria, 144
Ludgate, Nargiza, 68
- Masis, Tessa, 86
- Nejadgholi, Isar, 32
Nguyen, Dong, 1
- O'Connor, Brendan, 86
- Pacheco, Maria Leonor, 144
Pangakis, Nicholas J, 113
Pieuchon, Nicolas Audinet De, 57
Puri, Prateek, 68
- Rambow, Owen, 132
Roy, Shamik, 144
- Sadhu, Santadarshan, 68
Silcock, Emily, 99
Surdeanu, Mihai, 68
- Vallejo, Gisela, 16
Vasilets, Evgeny, 1
- Wang, Zhengxiang, 132
Wegmann, Anna, 1
Wolken, Sam, 113
- Yarlott, W. Victor H., 46