LREC-COLING 2024

The Second Workshop on Natural Language Processing for Political Sciences (PoliticalNLP 2024)

Proceedings

Editors:
Haithem Afli (General Chair), Houda Bouamor Cristina Blasi Casagran, and Sahar Ghannay

21 May, 2024
Torino, Italia

Proceedings of the The Second Workshop on Natural Language Processing for Political Sciences (PoliticalNLP 2024)

ISBN 978-2-493814-26-5
ISSN 2951-2093 (COLING); 2522-2686 (LREC)

Jointly organized by the ELRA Language Resources Association
and the International Committee on Computational Linguistics

Message from the General Chair

The PoliticalNLP 2024 workshop, marking its progression since its foundation, was set to be a focal point of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), convening on 21st May 2024. This event was poised to build upon the robust groundwork established by its inaugural session, which was conducted successfully on 24th June 2022 in Marseille, France, following the 13th Edition of LREC 2022. Initially concentrating on exploiting Natural Language Processing (NLP) techniques to extract, analyse, and interpret insights from socio-political data, the workshop dedicated itself to pioneering novel approaches in the text processing of socio-political content and their applications in information extraction and analysis.

Aiming to expand its thematic scope, this year's workshop invited contributions under the theme "Opportunities and Challenges of Generative AI and Large Language Models (LLMs) in Social and Political Sciences Research." It sought to encompass a broad array of NLP applications in socio-political contexts, striving to maintain its status as a premier forum for debating advanced language technologies within the social and political sciences. It encouraged collaborative dialogue among computational social and political scientists, computational linguists, machine learning practitioners, and researchers regarding the integration of NLP tools in contrast to conventional coding techniques.

PoliticalNLP 2024 aspired to continue this legacy by featuring distinguished speakers set to highlight contemporary topics relevant to the workshop's theme, facilitating a dynamic exchange of ideas on the potential benefits and challenges associated with generative AI and LLMs in socio-political research. Our heartfelt appreciation is extended to the Programme Committee members of the first workshop for their meticulous reviews and to the participants whose engaging discussions deepened the dialogue on the PoliticalNLP themes.

We also extend our profound gratitude to our sponsors, the Science Foundation Ireland Research Centre ADAPT (Grant Agreement No. 13/RC/2106_P2) and the UAB Research group on "GLOBAL security, technology and International Law," whose support was instrumental in making the second workshop a resounding success. As we advanced towards PoliticalNLP 2024, we looked forward with great enthusiasm to cultivating similarly enriching interactions, to advancing the dialogue on NLP applications within the realms of social and political sciences, and to exploring the continuously evolving field of generative AI and LLMs in this dynamic and vibrant domain.

Haithem Afli

Organizing Committee

General Chair:

Haithem Afli, ADAPT Centre, Munster Technological University, Ireland

Program Chairs:

Houda Bouamor, Carnegie Mellon University, Qatar

Sahar Ghannay, Paris-Saclay University, CNRS, LISN, France

Cristina Blasi Casagran, Autonomous University of Barcelona, Spain

Reviewers:

Wajdi Zaghouani, Hamad Bin Khalifa University, Qatar
Bruno Andrade, Munster Technological University, Ireland
Lenka Dražanová, European University Institute, Italy
Georgios Stavropoulos, The Centre for Research and Technology, Greece
Andrea Iana, University of Mannheim, Germany
Valentin Barrieren, Telecom ParisTech, France
Mohammed Hasanuzzaman, Munster Technological University, Ireland
Patrick Paroubek, Paris-Saclay University, CNRS, LISN, France
Suman Adhya, Indian Association for the Cultivation of Science, India
Valentin Barriere, Centro Nacional de Inteligencia Artificial, Santiago, Chile

Table of Contents

Workshop Program

Deciphering Political Entity Sentiment in News with Large Language Models: Zero-Shot and Few-Shot Strategies
Alapan Kuila and Sudeshna Sarkar

Event Detection in the Socio Political Domain
Emmanuel CARTIER and Hristo Tanev

Multi-Dimensional Insights: Annotated Dataset of Stance, Sentiment, and Emotion in Facebook Comments on Tunisia's July 25 Measures
Sanaa Laabar and Wajdi Zaghouani

Masking Explicit Pro-Con Expressions for Development of a Stance Classification Dataset on Assembly Minutes
Tomoyosi Akiba, Yuki Gato, Yasutomo Kimura, Yuzu Uchida and Keiichi Takamaru

Analysing Pathos in User-Generated Argumentative Text
Natalia Evgrafova, Veronique Hoste and Els Lefever

Knowledge Graph Representation for Political Information Sources
Tinatin Osmonova, Alexey Tikhonov and Ivan P. Yamshchikov

Analyzing Conflict Through Data: A Dataset on the Digital Framing of Sheikh Jarrah Evictions
Anatolii Shestakov and Wajdi Zaghouani

Semi-Automatic Topic Discovery and Classification for Epidemic Intelligence via Large Language Models
Federico Borazio, Danilo Croce, Giorgio Gambosi, Roberto Basili, Daniele Margiotta, Antonio Scaiella, Martina Del Manso, Daniele Petrone, Andrea Cannone, Alberto M. Urdiales, Chiara Sacco, Patrizio Pezzotti, Flavia Riccardo, Daniele Mipatrini, Federica Ferraro and Sobha Pilati

Towards quantifying politicization in foreign aid project reports
Sidi Wang, Gustav Eggers, Alexia de Roode Torres Georgiadis, Tuan Anh Đo, Léa Gontard, Ruth Carlitz and Jelke Bloem

Echo-chambers and Idea Labs: Communication Styles on Twitter
Aleksandra Sorokovikova, Michael Becker and Ivan P. Yamshchikov

Deciphering Political Entity Sentiment in News with Large Language Models: Zero-Shot and Few-Shot Strategies

Alapan Kuila, Sudeshna Sarkar
IIT Kharagpur
India
alapan.cse@iitkgp.ac.in, sudeshna@cse.iitkgp.ac.in

Abstract

Sentiment analysis plays a pivotal role in understanding public opinion, particularly in the political domain where the portrayal of entities in news articles influences public perception. In this paper, we investigate the effectiveness of Large Language Models (LLMs) in predicting entity-specific sentiment from political news articles. Leveraging zero-shot and few-shot strategies, we explore the capability of LLMs to discern sentiment towards political entities in news content. Employing a chain-of-thought (COT) approach augmented with rationale in few-shot in-context learning, we assess whether this method enhances sentiment prediction accuracy. Our evaluation on sentiment-labeled datasets demonstrates that LLMs, outperform fine-tuned BERT models in capturing entity-specific sentiment. We find that learning in-context significantly improves model performance, while the self-consistency mechanism enhances consistency in sentiment prediction. Despite the promising results, we observe inconsistencies in the effectiveness of the COT prompting method. Overall, our findings underscore the potential of LLMs in entity-centric sentiment analysis within the political news domain and highlight the importance of suitable prompting strategies and model architectures.

Keywords: zero-shot, few-shot, sentiment analysis, chain-of-thought prompting, in-context learning, self-consistency

1. Introduction

Sentiment analysis (SA) is a vital area in natural language processing (NLP)(Liu, 2020), focused on deciphering opinions and emotions using computational methods (Poria et al., 2020). It has diverse applications, from product reviews to social media insights. Previous research has addressed sentiment analysis at various levels, such as sentence, paragraph, and document levels (Zhang et al., 2023). Moreover, studies have focused on different targets of sentiment, including overall sentiment, aspect-based sentiment (Brun and Nikoulina, 2018), and sentiment associated with event mentions (Zhang et al., 2022). Analyzing sentiment pertinent to the salient entities in the news article is an important problem in computational journalism and news content analysis (Rønningstad et al., 2023). In the context of political natural language processing (NLP), understanding the sentiment towards political entities in news articles is particularly crucial. Political entities, such as countries, politicians, and political organizations, often drive the narrative in news coverage. Therefore, being able to accurately assess the sentiment towards these entities can provide valuable insights into public opinion, political discourse, and media framing.

Similar to works by (Tang et al., 2023), and (Bastan et al., 2020), our research specifically targets sentiment analysis related to particular entities. Historically, sentiment analysis relied on bag-of-word models, which failed to capture word ordering, a crucial aspect of sentiment prediction. Later, machine learning (ML) and deep learning (DL) models gained popularity for sentiment analysis tasks, though they struggled with generalization on domain-specific datasets (Kenyon-Dean et al., 2018). Recently, techniques like transfer learning (Golovanov et al., 2019) and self-supervised learning (Qian et al., 2023) have been applied to improve model generalization and reduce data dependence, particularly demonstrating promising performance in few-shot settings with limited annotated data. However, state-of-the-art deep neural network models remain complex and opaque in their decision-making processes, posing challenges for both end-users and system designers.

However, recent research on pre-trained large language models (LLMs) has demonstrated impressive performance across a variety of natural language processing (NLP) tasks, particularly in common sense reasoning (Brown et al., 2020). These LLMs have proven capable of generalizing to new tasks using zero-shot and few-shot learning, facilitated by suitable prompts and in-context learning (Huang and Chang, 2022). Moreover, the introduction of chain-of-thought (COT) prompting (Wei et al., 2022) has further enhanced the reasoning abilities of LLMs by generating intermediate reasoning steps. By incorporating rationale into the prompt design and providing (input, output) instance-pair demonstrations, the COT approach encourages LLMs to generate textual explanations alongside predicting the final output. Additionally, self-consistency mechanisms (Wang et al., 2022) reinforce the reasoning capabilities of LLMs

Proceedings of the Second Workshop on Natural Language Processing for Political Sciences (PoliticalNLP 2024), pages 1–11

through *sample-and-marginalize* decoding procedures. Despite these advancements, some studies have suggested that accumulating explanations with prompts during in-context learning may have adverse effects on LLM performance in question-answering (QA) and natural language inference (NLI) tasks (Ye and Durrett, 2022). Nevertheless, LLMs have proven effective in various textual reasoning tasks, including arithmetic and symbolic reasoning problems (Wang et al., 2022). In our research, we aim to investigate whether LLMs can accurately predict entity-centric sentiment polarity from political news text. By exploring the intersection of large language models and sentiment analysis, we hope to shed light on the capabilities and limitations of these models in understanding and interpreting sentiment dynamics in textual data.

To employ large language models (LLMs) for predicting entity-specific sentiment, we harness the chain-of-thought (COT) mechanism to guide prompt design. In our zero-shot chain-of-thought approach, we adopt a two-stage prompting strategy. Initially, we extract the contextual justification for the prediction, followed by returning the final sentiment label in the second stage. Our few-shot approach involves integrating a limited number of (entity context, entity-centric sentiment label, rationale) triplets into the LLMs during training. Here, the entity context may encompass a sentence, paragraph, or entire document, while the entity-centric sentiment label denotes the sentiment polarity towards the target entity as depicted in the context. The rationale comprises one or more sentences elucidating the reasoning behind the predicted outcome. We assess our model's performance based on the accuracy of the final predicted sentiment class.

Additionally, previous research has highlighted the necessity of scaling up LLMs with several hundred billion parameters, such as the OpenAI GPT series (GPT-3-175B) (Brown et al., 2020), PaLM(540B) (Chowdhery et al., 2022), and LaMDA(137B) (Thoppilan et al., 2022), to achieve satisfactory performance in COT scenarios. However, adopting these large pre-trained language models may be unfeasible for many users with resource constraints (Ranaldi and Freitas, 2024). In our experiments, we employ LLMs with relatively fewer model parameters, namely Mistral-7B (Jiang et al., 2023), LLaMA2-13B (Touvron et al., 2023) and Falcon-40B (Almazrouei et al., 2023). By deliberately altering various aspects of the demonstrated rationale and conducting a series of ablation experiments, we measure how the model's performance varies accordingly. Our extensive results demonstrate the effectiveness of LLMs with relatively fewer parameters in the task of entity-centric sentiment prediction from political news articles.

Our contributions are as follows:

- We explore the capability of LLMs to predict entity-specific sentiment from the news context in a zero-shot setting.
- We examine the efficacy of the Chain-of-Thought (COT) approach in conjunction with Large Language Models (LLMs), bolstered by rationale in few-shot in-context learning. Our objective is to determine whether this combined approach improves the model's ability to predict entity-specific sentiment from document-level context.
- We evaluate the accuracy and robustness of our proposed approach using two sentiment-labeled news datasets. The first dataset [1] comprises political news articles sourced from the Event-Registry API [2], while the second dataset [3] is obtained from (Bastan et al., 2020), providing diverse contexts for evaluation and comparison.

2. Entity Centric Sentiment from Political News

In this paper, we address the task of determining the overall sentiment polarity expressed towards a target entity in a political news article. This task differs significantly from existing works on sentiment prediction in movie reviews, product reviews, or social media datasets (Kumaresan and Thangaraju, 2023). Unlike review text or social media datasets, news articles contain descriptive content with a significant amount of redundant information that is often irrelevant for sentiment prediction of the target entity. Additionally, subjective opinions are sometimes presented as objective information, posing challenges for automatic classifiers. While it may be easy for humans to discern the inherent sentiment polarity, automatic classifiers face difficulties in extracting the target entity-specific context from news articles, especially when multiple entities are mentioned multiple times, both directly and indirectly (Fei et al., 2023).

Moreover, a single news article may contain multiple opinions directed towards the target entity, and the sentiment towards the same entity may vary across different paragraphs within the same article. Hence, navigating through irrelevant information to extract the target entity-specific context and predict the correct sentiment becomes challenging.

[1] https://github.com/alapanju/EntSent

[2] https://github.com/EventRegistry/event-registry-python

[3] https://github.com/StonyBrookNLP/PerSenT

In the following sections, we first describe our approach, followed by the experimental details, including the dataset used and the LLM models employed. Subsequently, we present our experimental findings and engage in pertinent discussions. Following this, we delve into a detailed examination of existing works within this domain. Finally, the paper concludes by summarizing key discoveries and outlining avenues for future research.

3. Our Approach

In this paper, we explore the natural language understanding capabilities of the LLMs by predicting the entity-specific sentiment label from news articles in zero-shot and few-shot settings.

3.1. Zero-shot approach

In zero-shot settings, we do not utilize any training exemplar for model supervision. In the absence of demonstration exemplars, the LLM is provided with the prompt containing the problem definition, input context, and sentiment class labels. Problem definition denotes the task name (i.e. sentiment classification); Input context contains the news text and the target entity name; The sentiment class labels contain the set of final sentiment tags. The prompt defines the expected structure of the output that eventually helps us to decode the LLM-responses into our desired format. In our experiment, we utilize two prompting techniques for zero-shot sentiment classification.

standard zero-shot In standard zero-shot setting, the input prompt contains the task definition, query text and target entity as described in the figure 1. Due to clarity and space constraint, we present a single sentence instead of the whole document as the input text in that figure.

2-stage Prompting To comprehend the sentiment towards a particular entity, it is essential to discern both implicit and explicit opinions within the news context concerning that entity. Recent studies, such as (Kojima et al., 2022), have demonstrated that a two-stage prompting strategy, referred to as zero-shot Chain of Thought (COT), can enhance the performance of Large Language Models (LLMs) in various reasoning tasks. Consequently, we adopt a similar two-stage prompting approach in our methodology. In the first stage, we extract textual cues indicating how the target entity is depicted sentiment-wise. Subsequently, in the second stage, we predict the final sentiment label. By employing this dual prompting process, we obtain the sentiment label pertaining to the target entity. Figure 2 illustrates the 2-stage prompting method, depicting both the intermediate and final outputs. The intermediate output serves as an explanation for the final sentiment prediction.

3.2. Few-shot approach

(Brown et al., 2020) have shown that LLMs can perform new tasks during inference when prompted with a few-demonstrations. In our experiment, we follow different prompting strategies and measure the LLM performance on few-shot scenario.

Standard few-shot In standard few-shot prompting, the LLMs are provided with in-context exemplars containing (input, output) pairs before providing the query input text. Here, the input is the entity context (news article) and the output is the entity-specific sentiment tag (positive, negative or neutral). The sample input prompt as well as the generated output is reported in the figure 3.

COT prompting In COT prompting, we augment the (input, output) example-pair demonstrations with a natural language rationale that demonstrates the justification of the output sentiment tag. Hence the prompt is a triplet containing (input, rationale, output). By incorporating rationale, we aim is to investigate whether LLMs can benefit from these explanations when learning from exemplars in-context. The sample input prompt is illustrated in figure 4.

3.3. Self-Consistency

Upon scrutinizing news articles, it becomes evident that the portrayal of the same entity varies across different paragraphs within a single article. Hence, to ascertain the overall document-level sentiment of the target entity, it is imperative to encompass and weigh all these opinions pertaining to the entity. (Wang et al., 2022), introduced the concept of self-consistency, which revolves around generating multiple reasoning paths to determine the correct final answer. Leveraging this concept, we aim to influence the decoder of Large Language Models (LLMs) to produce a diverse set of reasoning paths for predicting the final sentiment tag. Following (Wang et al., 2022), in the self-consistency method, we first prompt the LLM using chain-of-thought prompting. Subsequently, instead of employing a greedy search decoding approach, we utilize various existing sampling algorithms to generate a diverse set of candidate reasoning paths. Each of these paths may lead to a different final sentiment label. Finally, employing a majority voting approach, we marginalize the sampled reasoning paths and select the sentiment label that remains consistent across all generated

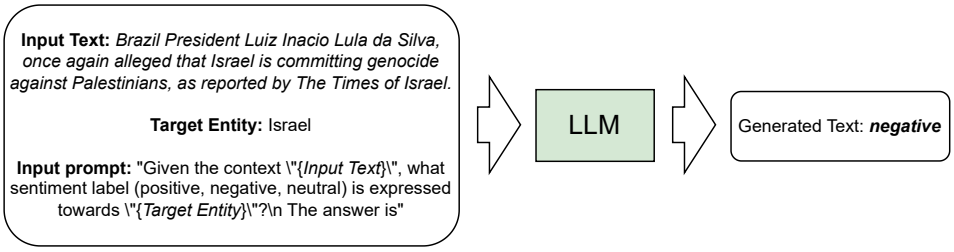


Figure 1: Standard Zero-shot Prompting Demonstration. Input prompt includes news article excerpt and target entity phrase, generating entity-specific sentiment output.

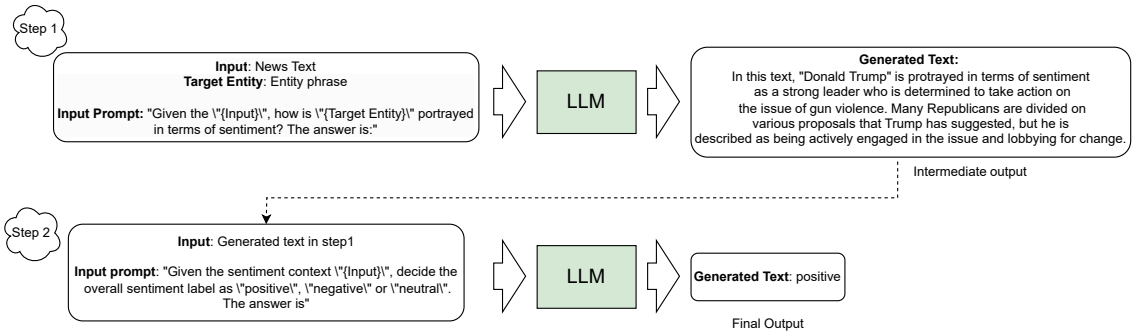


Figure 2: Two-Stage Prompting for Sentiment Prediction in Zero-Shot Setting. The first stage involves extracting rationales for entity-specific sentiment prediction, providing opinions regarding the target entity. In the second stage, sentiment tags are predicted based on the explanations

answers. Through this "sample-and-marginalize" decoding method, our objective is to encapsulate all sentiment-inducing components from the news content and amalgamate them to determine the overall sentiment. The self-consistency method is depicted in Figure 5.

4. Experimental Setup

4.1. Datasets

In our experiment, we utilize datasets released by (Bastan et al., 2020) for our model evaluation. Additionally, we curate a news dataset focused on the political domain, extracted and annotated by the Event-Registry API. These datasets consist of news articles, target entity phrases, and entity-specific sentiment tags, but lack rationale information. In the following subsections, we provide detailed descriptions of these two datasets.

PerSenT: The PerSenT dataset, introduced by (Bastan et al., 2020), is designed to predict the author's sentiment towards the main entities in news articles. This dataset includes paragraph-level as well as entire document-level sentiment annotations towards the target entity. The authors have partitioned the entire dataset into train, development, and test splits. Additionally, the paper reports the performance of various fine-tuned BERT model variants on this dataset. In our experiment, we evaluate the performance of our proposed LLM models on the test set of the PerSenT dataset.

Dataset	Positive	Neutral	Negative	Total	Unique Entites
PerSenT Test-Std	293	213	73	579	426
PerSenT Test-Freq	368	320	139	827	4
WPAN	600	600	600	1800	3

Table 1: Comparison of Test Dataset Statistics: PerSenT vs. WPAN for Entity-Specific Sentiment Analysis

Event Registry Data: The PerSenT dataset mentioned earlier contains a diverse selection of news articles from various domains. However, to explore how different countries and their policies are portrayed in the media, we have compiled a new dataset focused on global politics. This involved gathering news articles pertaining to specific nations and their policies from media outlets worldwide. We named the dataset as WPAN (**W**orldwide **P**erception **A**nalysis of **N**ations)[4]. The selected

[4] https://github.com/alapanju/EntSent

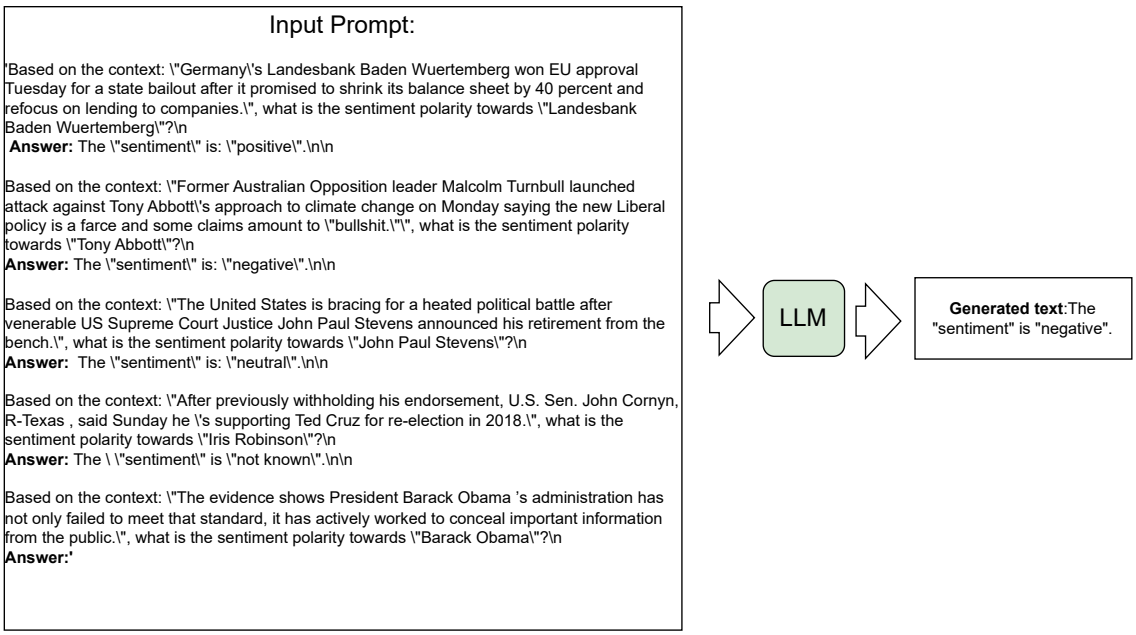


Figure 3: Standard Few-shot Prompting Illustration. Input: Entity context (news article). Output: Entity-specific sentiment tag (positive/negative/neutral).

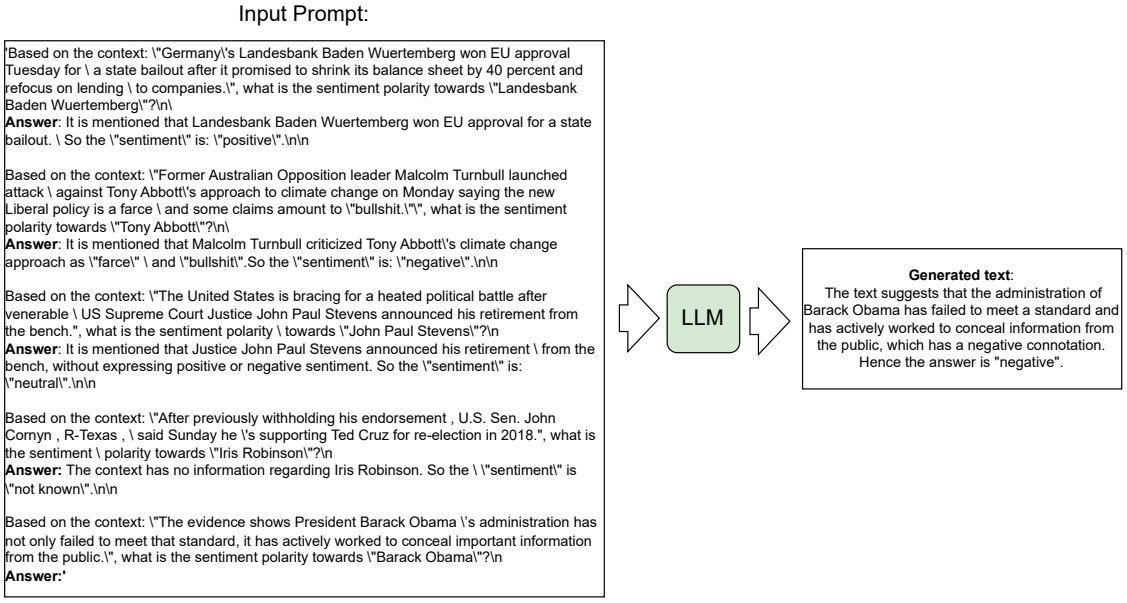


Figure 4: COT Prompting Demonstration. Input: Entity context (news article), Rationale: Justification of sentiment tag, Output: Entity-specific sentiment (positive/negative/neutral).

nations include India, Russia, and Israel. For India, we included media sources from neighboring countries such as Pakistan and Bangladesh. Similarly, for Russia, we selected outlets from India, the UK, and the USA, while for Israel, we included outlets from India and Pakistan. Our selection criteria were based on the significance of these nations in events such as the Russia-Ukraine conflict, the Israel-Hamas conflict, and tensions in the Indian subcontinent. We chose media sources based on article frequency, ensuring a random selection without bias. We used the Event-Registry Python API to extract relevant news articles. The API also provides a target specific document level sentiment score. The sentiment scores range from -1 to 1, where 1 represents maximum positive sentiment. We categorized sentiment scores between 0.6 to 1 as positive, -0.2 to $+0.2$ as neutral, and -0.6 to

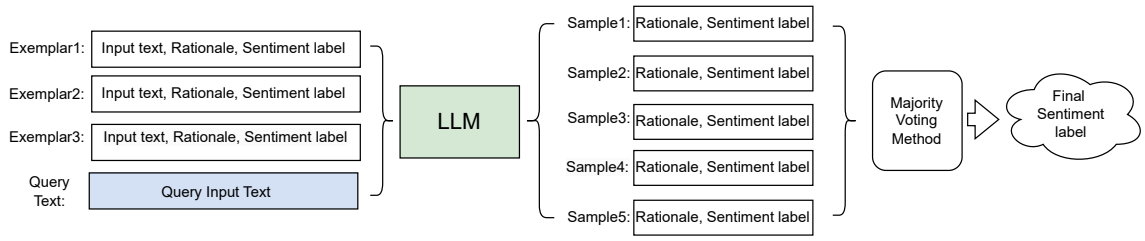


Figure 5: Chain-of-Thought Prompting with Self-Consistency in few-shot settings. The image illustrates the process where the LLM is provided with demonstration exemplars consisting of (input, output, and rationale) triplets. During inference, given a query input, the LLM returns multiple outputs containing sentiment tags and related explanations. The final sentiment tag is identified through sampling and marginalization technique.

-1 as negative. We collected 200 articles for each sentiment range for each target topic, resulting in 600 articles per topic. Each dataset record includes the news article, target entity, and entity-specific sentiment label.

4.2. Language Models

Our task integrates entity-specific sentiment analysis with elements of reasoning, particularly in justifying assigned sentiment based on contextual cues and linguistic patterns in the text. Traditional deep learning models or small language models (SLMs) require rationale-augmented training, which is costly and challenging to scale. However, in-context few-shot learning via prompting and the COT approach significantly enhance LLMs' reasoning capabilities across various tasks. Through COT, LLMs can perform few-shot prompting for reasoning tasks using triplets: (input, output, chain-of-thought). Studies have demonstrated the effectiveness of COT in improving reasoning abilities in LLMs with large parameter sizes, which are often inaccessible due to resource constraints. In this study, we aim to determine whether LLMs with fewer parameters can generate explicit reasoning while predicting entity-specific sentiment classes from document context. To explore this, we experiment with three transformer-based language models of varying scales:

- Mistral (Jiang et al., 2023) is a open-sourced decoder-based model with 7-billion parameters.
- Llama-2 (Touvron et al., 2023), developed by Meta AI, is a collection of transformer-based language models ranging in scale from 7 billion to 70 billion parameters. In our experiment, we use fine-tuned model named Llama2-13b-chat-hf with 13-billion parameters.
- Falcon (Almazrouei et al., 2023) is a causal decoder-only open-sourced language model. In our experiment, we use *instruct* version of the language model with two different parameter size namely, Falcon-7b-instruct (7-billion) and Falcon-40b-instruct (40-billion).

4.3. Prompting and Decoding Scheme

We discuss about the prompt designing in the section 3. For few-shot setting, we employ 3-4 exemplars in our experiment. These samples are chosen randomly from the training set of the PerSenT dataset. In COT prompting method, the exemplar pairs are augmented with manually composed natural language explanations. For a fair comparison, we use same prompt structure and same exemplar sets across all the LLMs.

For standard zero-shot and few-shot methods, we use greedy search decoding. In the case of 2-stage prompting method, we first use top-k, top-p and temperature sampling and in the second step, we employ greedy decoding method. In self-consistency method, instead of greedy search decoding, we employ different sampling algorithms like top-k sampling, temperature sampling, top-p sampling. In our experiment, the range of k value is between$\{50, 90\}$; The p value varies in the range of $\{0.9, 0.95\}$. We keep the temperature value as 0.7. We utilized a BERT model (Devlin et al., 2019) fine-tuned with the training data from the PerSent dataset as our baseline.

5. Result Analysis

In this section, we evaluate the performance of the Large Language Models (LLMs) based on the correctness of the final predicted sentiment labels using macro-F1 score metric for quantitative analysis. The experiments are conducted multiple times with different sets of training samples, and the average output over three runs is reported to ensure consis-

Model	Zero-shot			Few-shot		
	Std decoding	2-stage prompting	Self-Consistency	Std few-shot	COT prompting	Self-Consistency
Mistral-7b	42.16	43.32	45.67	49.87	49.56	52.78
Llama-13b-chat	41.59	42.92	43.21	49.43	50.88	51.97
Falcon-7b-instruct	41.47	42.81	44.22	48.56	49.87	52.63
Falcon-40b-instruct	43.89	44.13	47.05	50.24	51.39	54.94
Fine-tuned BERT	43.07					

Table 2: Macro F1-score for Document-level Entity-centric Sentiment prediction across Various LLMs on the PerSenT *Frequent* Test Dataset

Model	Zero-shot			Few-shot		
	Std decoding	2-stage prompting	Self-Consistency	Std few-shot	COT prompting	Self-Consistency
Mistral-7b	43.61	44.29	44.70	46.64	47.37	49.77
Llama-13b-chat	42.09	43.61	44.92	45.98	46.56	49.08
Falcon-7b-instruct	41.72	42.19	43.64	47.16	48.07	50.45
Falcon-40b-instruct	44.21	45.27	46.17	49.67	51.19	53.71
Fine-tuned BERT	48.38					

Table 3: Macro F1-score for Document-level Entity-centric Sentiment prediction across Various LLMs on the PerSenT *Standard* Test Dataset

Model	Zero-shot			Few-shot		
	Std decoding	2-stage prompting	Self-Consistency	Std few-shot	COT prompting	Self-Consistency
Mistral-7b	56.67	58.12	59.76	59.07	59.36	61.49
Llama-13b-chat	54.43	57.07	58.21	58.34	58.42	59.79
Falcon-7b-instruct	55.71	57.18	58.23	59.69	59.35	61.48
Falcon-40b-instruct	57.95	59.01	59.91	61.84	62.07	63.87
Fine-tuned BERT	56.54					

Table 4: Macro F1-score for Document-level Entity-centric Sentiment prediction across Various LLMs on the WPAN Dataset

tency. The seed value is fixed during experiments to obtain identical outputs.

Our analysis reveals variations in LLM performance between datasets. Specifically, the LLMs perform better on the WPAN dataset compared to the PerSenT dataset. Upon examining the news articles, we observed that most documents in the PerSenT dataset exhibit mixed sentiment across various paragraphs within the same article. In contrast, the sentiment across paragraphs in the WPAN dataset is less varied, potentially contributing to the improved performance of LLMs on this dataset.

We address three main experimental questions during the result analysis:

- We investigate whether LLMs can predict entity-specific sentiment labels in a zero-shot setting.
- We explore whether LLMs can learn from few-shot demonstrations.
- We analyze whether scaling up the LLM size has any effect on zero-shot and few-shot settings.

The experimental results, presented in Tables 2, 3, and 4, reveal several key insights.

Firstly, for the PerSenT-freq and WPAN datasets, the FALCON-40b model consistently outperforms the fine-tuned BERT model. Even in the case of PerSenT-std data, the Self-consistency method over FALCON-40b yields comparable performance compared to FT-BERT. These findings indicate that LLMs, with their pretraining and proper parameter sizing, exhibit a strong capability to capture sentiment labels from documents in zero-shot settings.

Secondly, we observe that the model performance improves significantly in few-shot settings compared to zero-shot scenarios across all three datasets. This suggests that learning in-context positively impacts model performance and effectiveness.

However, our experiments also reveal that the Chain-of-Thought (COT) prompting method is not consistently effective. In some cases, its perfor-

mance lags behind standard few-shot approaches. Nevertheless, the self-consistency method proves to be beneficial in enhancing model performance across all cases. Additionally, in zero-shot scenarios, the 2-stage prompting approach outperforms the standard zero-shot method.

Lastly, we experimented with LLMs having a parameter size within 40-billion. However, at this scale, we did not observe significant effects of model scaling on performance. Further exploration with larger model sizes may provide additional insights into this aspect.

Overall, our results demonstrate the effectiveness of LLMs in entity-specific sentiment prediction, particularly in few-shot learning scenarios, while highlighting the importance of appropriate prompting strategies and model architectures.

6. Related Work

6.1. Sentiment Analysis in news domain

Sentiment analysis is a widely explored area within natural language processing, attracting significant attention due to its multitude of applications across academic research and practical domains. Particularly within the realm of news content analysis, sentiment analysis has emerged as a pivotal task (Balahur et al., 2013; Katayama et al., 2019; Islam et al., 2017; Kuila et al., 2024; Samonte, 2018; Pryzant et al., 2019). Researchers have also delved into sentiment prediction grounded in news events (Zhou et al., 2021). Moreover, there is a burgeoning interest in news bias analysis (Eberl et al., 2017), which often relies on sentiment associated with news publications (Rodrigo-Ginés et al., 2024).

However, our focus in this paper lies specifically on entity-specific sentiment analysis within news articles. This presents a distinct problem with diverse applications, including predicting authors' sentiment (Bastan et al., 2020), discerning the ideology of news outlets (Lin et al., 2011), and analyzing media bias (Hamborg et al., 2019).

6.2. Large Language Models

Recent advancements in natural language processing (NLP) research have been marked by the emergence of Large Language Models (LLMs) (Chowdhery et al., 2022). These LLMs are pre-trained on massive text corpora using diverse training techniques such as instruction-tuning and reinforcement learning with human feedback (RLHF) (Christiano et al., 2017), showcasing impressive performance in zero-shot and few-shot settings. The paradigm shift towards in-context learning (Brown et al., 2020) has further enhanced the capabilities of LLMs, moving away from fine-tuning to prompt-tuning approaches.

Despite the widespread adoption of LLMs in various NLP tasks, including sentiment analysis (Zhong et al., 2023; Wang et al., 2023), challenges persist due to the computational demands of these models, particularly in resource-constrained settings. Consequently, our research is dedicated to investigating the viability of employing smaller-scale LLMs for entity-specific sentiment identification within the domain of political news articles. By harnessing the power of these compact LLMs, we endeavor to address resource limitations while leveraging the inherent capabilities of LLMs in sentiment analysis tasks.

7. Conclusion

In this study, we investigated the application of Large Language Models (LLMs) in predicting entity-specific sentiment from political news articles using zero-shot and few-shot strategies. Our findings demonstrate the effectiveness of LLMs, particularly FALCON-40b, in capturing sentiment towards political entities. Leveraging the chain-of-thought (COT) approach with rationale in few-shot in-context learning, we observed improvements in sentiment prediction accuracy, especially in few-shot scenarios. While the self-consistency mechanism enhanced consistency in sentiment prediction, we noted varying effectiveness in the COT prompting method across different datasets. Overall, our results highlight the potential of LLMs in entity-centric sentiment analysis within the political news domain.

Beyond sentiment analysis, our work has broader implications for media bias analysis and identification of media house ideologies. By discerning sentiment towards political entities, our model can assist in analyzing media bias and understanding the ideological stance of media houses. This capability holds promise for enhancing media literacy and facilitating informed discourse in political communication.

Moving forward, future research could explore additional applications of LLMs in political NLP tasks, such as misinformation detection, stance classification, and agenda setting analysis. Additionally, investigating the interpretability of LLMs' predictions and addressing potential biases in training data are essential considerations for further advancement in this field.

In conclusion, our study contributes to advancing the understanding of sentiment analysis in the political news domain and underscores the potential of LLMs in facilitating nuanced analysis of media content and political discourse.

8. Bibliographical References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra-Aimée Cojocaru, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models. *ArXiv*, abs/2311.16867.

Alexandra Balahur, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik Van Der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. 2013. Sentiment analysis in the news. *arXiv preprint arXiv:1309.6202*.

Mohaddeseh Bastan, Mahnaz Koupaee, Youngseo Son, Richard Sicoli, and Niranjan Balasubramanian. 2020. Author's sentiment prediction. *arXiv preprint arXiv:2011.06128*.

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.

Caroline Brun and Vassilina Nikoulina. 2018. Aspect based sentiment analysis into the wild. In *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 116–122.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Oliveira Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Paul Francis Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *ArXiv*, abs/1706.03741.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Jakob-Moritz Eberl, Hajo G. Boomgaarden, and Markus Wagner. 2017. One bias fits all? three types of media bias and their effects on party preferences. *Communication Research*, 44(8):1125–1148.

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat seng Chua. 2023. Reasoning implicit sentiment with chain-of-thought prompting. In *Annual Meeting of the Association for Computational Linguistics*.

Sergey Golovanov, Rauf Kurbanov, Sergey Nikolenko, Kyryl Truskovskyi, Alexander Tselousov, and Thomas Wolf. 2019. Large-scale transfer learning for natural language generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6053–6058, Florence, Italy. Association for Computational Linguistics.

Felix Hamborg, Anastasia Zhukova, and Bela Gipp. 2019. Automated identification of media bias by word choice and labeling in news articles. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 196–205.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Muhammad Usama Islam, Faisal Bin Ashraf, Ali Imam Abir, and M. Abdul Mottalib. 2017. Polarity detection of online news articles based on sentence structure and dynamic dictionary. *2017 20th International Conference of Computer and Information Technology (ICCIT)*, pages 1–5.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Daisuke Katayama, Yasunobu Kino, and Kazuhiko Tsuda. 2019. A method of sentiment polarity

identification in financial news using deep learning. *Procedia Computer Science*, 159:1287–1294. Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 23rd International Conference KES2019.

Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, Roman Sarrazingendron, Rohit Verma, and Derek Ruths. 2018. Sentiment analysis: It's complicated! 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 1886–1895, New Orleans, Louisiana. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *ArXiv*, abs/2205.11916.

Alapan Kuila, Somnath Jena, Sudeshna Sarkar, and Partha Pratim Chakrabarti. 2024. Analyzing sentiment polarity reduction in news presentation through contextual perturbation and large language models. *arXiv preprint arXiv:2402.02145*.

C. Kumaresan and P. Thangaraju. 2023. Elsa: Ensemble learning based sentiment analysis for diversified text. *Measurement: Sensors*, 25:100663.

Y. Lin, James P. Bagrow, and David M. J. Lazer. 2011. More voices than ever? quantifying media bias in networks. *ArXiv*, abs/1111.1227.

Bing Liu. 2020. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. 2020. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Transactions on Affective Computing*, 14:108–132.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2019. Automatically neutralizing subjective bias in text. *ArXiv*, abs/1911.09709.

Fan Qian, Jiqing Han, Yongjun He, Tieran Zheng, and Guibin Zheng. 2023. Sentiment knowledge enhanced self-supervised learning for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12966–12978, Toronto, Canada. Association for Computational Linguistics.

Leonardo Ranaldi and Andre Freitas. 2024. Aligning large and small language models via chain-of-thought reasoning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1812–1827, St. Julian's, Malta. Association for Computational Linguistics.

Francisco-Javier Rodrigo-Ginés, Jorge Carrillo de Albornoz, and Laura Plaza. 2024. A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it. *Expert Systems with Applications*, 237:121641.

Egil Rønningstad, Erik Velldal, and Lilja Øvrelid. 2023. Entity-level sentiment analysis (elsa): An exploratory task survey. *arXiv preprint arXiv:2304.14241*.

Mary Jane C. Samonte. 2018. Polarity analysis of editorial articles towards fake news detection. In *Proceedings of the 2018 1st International Conference on Internet and E-Business*, ICIEB '18, page 108–112, New York, NY, USA. Association for Computing Machinery.

Yixuan Tang, Yi Yang, Allen Huang, Andy Tam, and Justin Tang. 2023. FinEntity: Entity-level sentiment classification for financial texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15465–15471, Singapore. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam M. Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, Yaguang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, I. A. Krivokon, Willard James Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Hartz Søraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Díaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, V. O. Kuzmina, Joseph Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Rogers Croak, Ed Huai hsin Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *ArXiv*, abs/2201.08239.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava,

Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Huai hsin Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *ArXiv*, abs/2203.11171.

Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023. Is chatgpt a good sentiment analyzer? a preliminary study. *ArXiv*, abs/2304.04339.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *Neural Information Processing Systems*.

Qi Zhang, Jie Zhou, Qin Chen, Qingchun Bai, and Liang He. 2022. Enhancing event-level sentiment analysis with structured arguments. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1944–1949.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment analysis in the era of large language models: A reality check.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *ArXiv*, abs/2302.10198.

Deyu Zhou, Jianan Wang, Linhai Zhang, and Yulan He. 2021. Implicit sentiment analysis with event-centered text representation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6884–6893, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Event Detection in the Socio Political Domain

Emmanuel Cartier, Hristo Tanev
European Commission, Joint Research Center
Via Enrico Fermi, 2749
21027 Ispra (VA), Italy
emmanuel.cartier@ec.europa.eu, hristo.tanev@ec.europa.eu

Abstract

In this paper we present two approaches for detection of socio political events: the first is based on manually crafted keyword combinations, and is implemented inside the NEXUS event extraction system, and the second one is based on a BERT classifier. We compare the performance of the two systems on a dataset of socio-political events. We also evaluated only NEXUS on the ACLED event dataset, in order to show the effects of taxonomy mapping and the performance of rule based approaches. Interestingly, both systems demonstrate complementary performance. Both showing their best performance on different event type sets. Nevertheless, an LLM data augmented dataset shows that in this case the transformer-based system improves considerably. We also review in the related work section the most important resources and approaches for event extraction in the recent years.

1. Introduction

1.1. NEXUS event taxonomy

Event extraction started to emerge as a Computational linguistics topic of interest, in relation to the enormous stream of events reported in mainstream media and commented and repeated in the social networks (Kounadi et al., 2015). Event extraction is used in a wide range of applications in diverse domains and has been intensively researched for more than three decades, starting with the seminal works, inspired by the Message Understanding Conferences (Chinchor and Marsh, 1998). It has a large range of applications in policy making, security, disaster management, health, bio medical research, as well as in the domain of business and finances.

In recent years, the significance of event extraction in the socio political domain has garnered considerable attention from the research community. This heightened interest stems from the critical nature of socio-political phenomena and the escalating societal and political tensions witnessed over the past half-decade, attributed to events such as the COVID-19 pandemic, the conflict between Russia and Ukraine, and various other theatres of war, notably in the Middle East. The significance of event extraction technology in the socio-political realm has been underscored in recent workshops such as the CASE (Challenges and Application of Automated Extraction of Socio-political Events) series (Hürriyetoğlu et al., 2021) and other similar venues.

The purpose of this paper is to throw light on the most important approaches and resources for event extraction in the last years, illustrating the two predominant paradigms for event detection: the rule based and the statistical one by evaluating two event detection systems.

The statistical system is based on XLM RoBERTa-base statistical classifiers and the rule-based system Tanev et al. (2008), NEXUS, uses a set of manually curated boolean combinations of keywords. We compare the performance of the two systems on the JRC corpus of crisis events (Atkinson et al., 2017a). We additionally evaluate NEXUS on a subset of the ACLED dataset, which is a standard in the socio-political field, (Raleigh et al., 2010). The purpose of this evaluation was to study how well the NEXUS event types map to the ACLED taxonomy and to explore the effect of taxonomy alignment in the evaluation of event extraction systems.

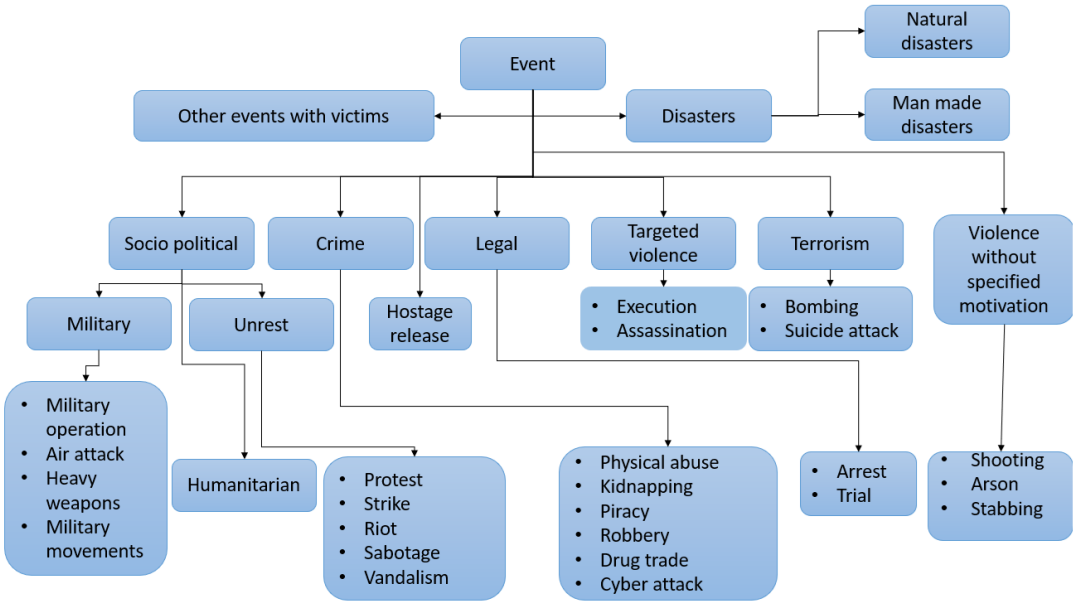
2. Related work

Rule-based event extraction was a predominant paradigm in the early systems in the nineties , as well in the next decade Aone and Ramos-Santacruz (2000); Grishman et al. (2002b,a). However, with the advent of the "big data" paradigm, state-of-the-art research experiments nearly entirely shifted towards the domain of Machine Learning (ML) and Large Language Models (LLM) (Hürriyetoğlu et al., 2021). Nevertheless, rule based systems have been dominating the industrial landscape (Chiticariu et al., 2013) and still provide basis for event detection in the domain of security and media analysis Tanev et al. (2008); Nitschke et al. (2022); Hamborg et al. (2019).

Building machine learning models for event detection was greatly facilitated by the availability of annotated event corpora and event databases. Among the known event corpora, one of the most used one is the ACE corpus (Consortium et al., 2005). Recently, the Joint Research Centre of the

Proceedings of the Second Workshop on Natural Language Processing for Political Sciences (PoliticalNLP 2024), pages 12–21

Figure 1: NEXUS event taxonomy



European Commission proposed two corpora, one of them based on the output of the NEXUS event extraction system, (Atkinson et al., 2017b), and the other one containing events related to the COVID pandemic (Piskorski et al., 2023). But this second one has another ontology than the NEXUs system and won't be used.

Security-related event databases (DB) are manually curated, such as ACLED (Raleigh et al., 2010) or automatically created, such as GDELT (Ward et al., 2013). Each DB record describes a security event with its time, location, event type, main actors, their nationality, victims, and optionally a text describing the event. Other well known socio-political databases are POLDEM (Kriesi et al., 2020), POLECAT (Halterman et al., 2023), UCDP data set (Sundberg et al., 2012). An overview of the publicly available event databases is provided in (Olsen et al., 2024).

3. EMM NEXUS

This section briefly describes the real-time event extraction system NEXUS (News cluster Event eXtraction Using language Structures). It is a rule based system, which uses Boolean combinations of keywords for event detection and grammar rules for event argument extraction.

NEXUS is an integral part of the Europe Media Monitor (EMM) and it has been described in details in (Tanev et al., 2008) and (Tanev et al., 2009). Its event taxonomy, see Figure 1 has been used to create the JRC security event corpus, described in (Atkinson et al., 2017a). Moreover, the corpus was created by manually annotating and curating articles, with events pre-detected by the system.

NEXUS uses the clusters of news articles, created by the EMM software (Tanev et al., 2008). Clusters describing various types of crisis events are selected via application of combinations of keywords, manually crafted and expanded with the help of terminology extraction software. The NEXUS system detects and extracts one main crisis event for each news cluster reporting an event of interest.

For each event the system generates a frame, whose main slots are: date and location, number of killed and injured, kidnapped people, actors, and type of event.

Noteworthy, NEXUS processes only the title and three leading sentences for each news article in the news article cluster and then it fuses the event information, extracted from the different articles. The system uses finite state cascade grammar rules over dictionaries of linear grammar patterns **<person> was found dead** or **<person> was stoned to death**. The semi-automatic learning of these patterns and the accompanying lexicon with references to names, professions, organizations, numbers, and other entities, were described in (Tanev et al., 2009).

Event types are detected via a set of keyword based boolean rules. In Table 1 we show excerpts from such rules for the event types *armed conflict*, *riot*, and *air attack*. It is important to consider the following: When processing clusters of news articles, keywords are searched in the title and in the first three sentences of each article in the clus-

Table 1: Samples from the Boolean keyword combinations for event detection

Type	Rule	
riot	AND	"hundreds of angry" OR "demonstration against" OR "mutiny" ...
		"clashes" OR "clashed" OR "burnt" OR "torched" OR "disperse" ...
armed conflict	AND	"troops" OR "soldiers" OR "rebels" OR "insurgents" ...
		"deployed" OR "clashed" OR "battling" OR "returned fire" ...
	AND	"marines" OR "armed forces" OR "troops" ...
		"militants" OR "insurgents" OR "rebels" ...
air attack	AND	"fighter plane" OR "jets" OR "missile" OR "gunship" OR "interceptor" ...
		"damaged" OR "intercepted" OR "pounded" OR "targeted" ...
	-	"helicopter fired" OR "air raid" OR "missile attack" OR "bombing run" ...

ter. Second, each keyword combination has an assigned maximal word proximity. For example, considering the air attack keyword combination, its proximity is defined to be 17 tokens. Consequently, if both the word "jets" and "intercepted" appear in no more than 17 tokens from each other in the first 3 sentences of a news article, the "air attack" event will be triggered.

For several event types, NEXUS requires not only keyword rules to fire, but also the presence of dead or injured victims, detected by the argument extraction grammars. This serves as an additional filter, which increases the precision.

The event type taxonomy, recognized by the NEXUS system reflects the requirements of the Joint Research Centre's Europe Media Monitor (EMM). The event types, recognized by NEXUS are the most frequently reported in the news event classes, referring to crises.

The recognized crisis event types encompass a subset of the security and socio-political events, reported in the news, including man made incidents and natural disasters. Figure 1 shows the taxonomy of the event types, which are a focus of the system. These events can be grouped into several large classes:

1. Socio-political events: They encompass all unrests, protests, military operations, and humanitarian crises. The "unrest" subtypes include violent unrests like *riots*, but also*protests*, *strikes and boycotts*, as well as *sabotages*. Military events involve *armed conflicts*, i.e. battles and sieges performed by military and organized armed groups, *air and missile attacks*, as well as *exploitation of heavy weapons* such as artillery and heavy firing arms. Military events include also *deployment and movements of troops and military vehicles*. Humanitarian events include reports about displacement of people and lack of resources, such as food, water, shelter, and medicines.

2. Crimes: NEXUS recognizes *kidnapping*, *robbery*, *pirate attacks on ships*, *physical abuse*, and *cyber attacks*. Physical abuse events include also cases of sexual abuse. In reality crimes can be part of a terrorist operation, for example kidnapping of a political leader. Similarly, cyber attacks are used as a unconventional warfare and in some cases can also be classified as terrorist attacks. However, given the multifaceted nature of these event classes, they are put in the crime category both for simplicity in classification, as well as because their nature is related to the violation of the law.

3. Legal events: The system detects two legal event types, which are related to the security, namely *arrests* and *trials*.

4. Targeted violence: These are violent events, who are directed towards pre-defined people. The term "targeted violence" is taken from the PLOVER socio political event ontology (Halterman et al., 2023). According to this ontology two event types are considered as targeted violence, namely *execution* and *assassination*.

5. Terrorist attacks: NEXUS recognizes the most common forms of terrorist attacks: namely bombings, including suicide attacks, as well as all violence, explicitly labeled as terrorism or performed by certain armed groups (e.g. Al Qaeda, IRA, etc.).

6. Violence without detected motivation: The three event types of *shooting*, *stabbing*, and *arson* fall in this category, when the system cannot detect the motivation context, which could be crime, terrorism, unrest, or military.

4. Experiments and Evaluation

4.1. Evaluating NEXUS

4.1.1. Evaluation on the JRC event corpus

The NEXUS system has been used, when creating the JRC security event corpus (Piskorski et al., 2023). First, the events were detected by NEXUS,

and then they were manually moderated and errors were corrected. The taxonomy of NEXUS was used when labeling the events from the JRC security corpus.

The JRC corpus contains around 617K events, extracted by NEXUS, of which 17K are manually curated. The authors of the corpus provided also a detailed evaluation of the event type, geolocation, and argument detection accuracy of the NEXUS event detection system. However, they use a very limited gold standard of 16 news clusters. In contrast, we wanted to evaluate the event classification of NEXUS on a proper subset of the manually moderated JRC corpus. In this paper we report on evaluation of the English language part of the manually moderated part of the corpus, which contains 7,934 detected events, each provided with a manually selected event type code, a title, and a text fragment, containing one or two sentences describing the event.

We have run NEXUS on the title and the event describing fragment in each of the 7,934 English language events from the corpus and compared the extracted event types against the ground truth annotation. Then, we have measured the precision, recall and F1 measure. Results are reported in Table 2.

Clearly, the NEXUS system works best for the "Unrest" event type among all socio-political events. The unrest involves all the protests, riots, and violent anti government actions, which are not terrorism. The legal event types, "Arrest" and "Trial" are also among the best performing classes. It was disappointing the low results for the military event types. Notably, we have got very low recall for all the event types "Military operation", "Air attack", and "Heavy weapons". These low results in the military event types clear suggest how to further improve the system.

The system works quite well also on the disaster group of event classes.

4.1.2. Preliminary Evaluation on the ACLED event database

We have conducted an additional evaluation of the NEXUS system on a more standard and widely used event data set.

For this purpose we have chosen the ACLED event dataset (Raleigh et al., 2010). It is one of the largest manually curated event databases. Evaluating against ACLED however was related to the challenge of of mapping NEXUS event types to the ACLED ones.

There are some little differences of the definitions of the event types of ACLED and NEXUS: first, ACLED does not cover incidents and disasters. Second, it does not classify explicitly events as terrorist attacks, but puts part of them in the larger category of "Explosions/Remote violence". Moreover ACLED events encompass also peaceful events, called "Strategic developments" and in NEXUS only one event type, namely "Arrest" is included in this class. Independently of these differences, we have managed to map some of the NEXUS event types into ACLED categories. Mapping was most of the time many to one: many NEXUS categories were mapped to one ACLED class. In Table 3 we show the mapping between the two event classification systems. In Table 4 we report the results from the ACLED evaluation after the mapping took place. What is important is that first, we cover only a small percent of the strategic developments; second, we did not manage to map properly terrorist events, since they are not part of the ACLED taxonomy and they were considered like no events. Therefore, the ACLED evaluation we performed can be considered approximate.

Still, the relations between the performance on different event types show similar trends in both evaluations: The "Protest" event type, which is a subtype of "Unrest", has a relatively high performance in the ACLED evaluation, as its super type "Unrest" has a good performance in the JRC corpus evaluation. Moreover, the system obtains low recall and low F1 score on the ACLED "Battle event", and similarly its corresponding NEXUS "Military operation" shows the same trend in the JRC corpus evaluation. Also, the ACLED Explosion/Remote violence which corresponds to the NEXUS "Heavy weapons", "Air attack" and "Bombing" obtains low recall, as its corresponding NEXUS types in the JRC corpus evaluation.

The conclusions drawn from both evaluations indicate that mapping between event taxonomies poses challenges, such as: partially overlapping event types, one to many event type relations, taxonomy gaps (for example the lack of terrorist attack in ACLED). The evaluation we have conducted on the ACLED data demonstrate these challenges.

On the other hand, this evaluation was also useful, since it confirmed several trends, observed in the JRC corpus evaluation, namely a notable underperformance of event detection rules in identifying military events and relatively high accuracy in modeling "Unrest" and its subtype "Protest".

4.2. Comparing Nexus to a Transformer-based system

So as to assess the respective merits of rule-based and transformer based systems, we fine-tuned a XLM-Roberta-base system on the JRC corpus. As this kind of system is sensible to dataset balance, we first give some general figures on the JRC corpus. Figure 2 shows the unbalanceness of this dataset of 6,892 annotated sentences.

Table 2: Performance of NEXUS on the JRC corpus

Event Type (code)	**Precision**	**Recall**	**F1**
Socio political			
Military operation (ARM)	0.66667	0.25586	0.36979
Air/missile attack (AA)	0.81395	0.30702	0.44586
Heavy weapons (HW)	0.52000	0.36111	0.42623
Terrorist Attack (TA)	0.63071	0.74146	0.68161
Bombing (BO)	0.67164	0.60811	0.63830
Unrest (SP)	0.83877	0.77140	0.80368
Humanitarian (HUM)	0.51485	0.4000	0.44835
Legal			
Arrest (AR)	0.92012	0.62854	0.74688
Trial (TRIAL)	0.92181	0.38063	0.53879
Crimes			
Kidnapping (KD)	0.73810	0.70992	0.72374
Physical abuse (PA)	0.55556	0.31746	0.40404
Non violent			
Hostage Release (RE)	0.83721	0.39560	0.53731
Violence without defined motivation			
Shooting (SH)	0.84834	0.47733	0.61092
Stabbing (ST)	0.73171	0.58824	0.65217
Targeted killing			
Execution (EX)	0.76190	0.64000	0.69565
Accidents and Disasters			
Earthquake (EQ)	0.90278	0.67708	0.77381
Flood (FL)	0.77477	0.74783	0.76106
Winter storm (IR)	1.00000	0.71875	0.83636
Storm (SR)	0.81481	0.62857	0.70968
Tropical storm (TR)	0.84211	0.68571	0.75591
Wild fire (WF)	0.96154	0.71429	0.81967
Landslides (LS)	0.73333	0.47826	0.57895
Man made disaster (MM)	0.86826	0.68289	0.76450
Maritime accident (MT)	0.94000	0.66197	0.77686
Explosion (XP)	0.68519	0.48684	0.56923
Another event type with dead or injured			
Other (NONE)	0.11632	0.75	0.20141
Accuracy			0.56479
Macro Avg	0.691476	0.61701	0.65035

Table 3: Mapping NEXUS to ACLED event types

ACLED category	**ACLED Explained**	**NEXUS**
Protest	Protests which start as peaceful	Protest
Riot	Riot or Mob Violence	Riot
Battle	Battle between organized forces	Military operation
Explosion/Remote violence	Bombings, shellings, air raids	Air Attack; Heavy Weapons ; Bombing; Suicide Attack
Strategic developments	Arrests, agreements, transfer of territories	Arrest
Violence against civilians	Violence against civilians	Physical Attack; Kidnapping

Table 4: Performance of NEXUS on the ACLED corpus

Class	Precision	Recall	F1-score
Battle	0.5995	0.3334	0.4285
Explosion/Remote violence	0.9356	0.2558	0.4018
Protest	0.8709	0.7278	0.7929
Riot	0.6607	0.1109	0.1899
Strategic developments	0.1794	0.3102	0.2273
Violence against civilians	0.8253	0.0695	0.1282
Accuracy			0.4153
Macro Avg	0.5816	0.2582	0.3098
Weighted Avg	0.7856	0.4153	0.4978

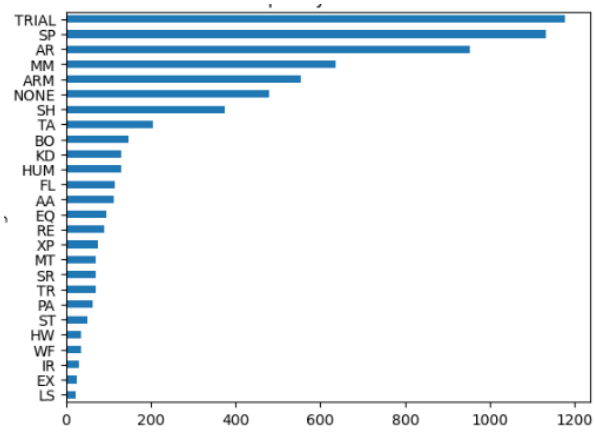


Figure 2: Distribution of classes in JRC news Dataset

We divided this dataset into the traditional training-development-test split: 80-10-10. it was done with the dataset Huggingface module thus respecting the distribution of the overall dataset. We train the model for 15 epochs, learning rate 2e-5, batch size 16. Table 5 shows the performance XLM Roberta achieves on the test set of EMM News. The overall accuracy is 0.67 after 15 epochs (after 8 epochs, we reached 0.70) surpassing the rule-based model by a large margin, especially when enough learning data are available. On this respect, the motto "more data, better results" is easily confirmed, pushing us to augment the EMM data.

4.3. Evaluating a Transformer-based system with augmented data

The last experiment we undertook consisted in balancing the fine-tuning dataset. Among several techniques to do so (REF), we decided to use LLM data augmentation techniques, as generative Language models reveal to be quite efficient in reformulating sentences (see e.g. (**?**)). As seen in the previous experiment, fine-tuning a model requires a lower-bound number of examples. To balance the dataset for under-represented classes, we used the following prompt:

> Your task is to generate {number} sentences, denoting the following type of event: {label}. As a help, the following sentence denote this type of event. To generate these sentences, please try to mimic a headline style, describing the facts and circumstances of the event. Generate these sentences in English, and rephrase the original sentence with several techniques, like synonym substitution, adverb insertion, paraphrasing and other distributional operations enabling to preserve the overall meaning while changing wording and phrasing. As output, please generate the sentences one per line. Be the most concise you can. Example sentence: {sentence}

where {number} represents the number of sentences to generate for every given class source example, calculated by the number of examples of the most represented class (Trial, 1,177) divided by the number of examples of the given class, rounded to the ceil. For example, for the class Arrest, 2 sentences will be generated for each source example (1,177 / 953 = 1.23 == 2); {label} represents the class, e.g. Arrest, and {sentence} represents the given example to rephrase, eg. *Nine held in Eta anti-terror raids*. Table 6 shows a few examples of paraphrases generated by GPT4 (OpenAI's June version with a context length of 8,192 tokens).

Figure 3 gives the distribution of samples per classes after data augmentation with a total of 35,583 example titles and on average more than 1,250 examples per class.

We then fine-tuned, with the same parameters as in the previous experiment, a language model. Table 7 shows the results on the EMM source titles for the sake of comparison with the Nexus system. As can be seen, the results are very promising, even if they need to be further confirmed on a totally new dataset. We also performed an error analysis, from the dispersion plot fig:displot.augmented.

Table 5: Evaluation results on JRC news dataset, fine-tuned XLM-Roberta-base model

event category	precision	recall	f1-score	support
Military operation (ARM)	0.52830	0.5	0.51376	56
Air/missile attack (AA)	0.64285	0.75	0.69230	12
Heavy weapons (HW)	0.0	0.0	0.0	3
Terrorist Attack (TA)	0.48	0.57142	0.52173	21
Bombing (BO)	0.33333	0.133333	0.190477	15
Unrest (SP)	0.75862	0.77192	0.76521	114
Humanitarian(HUM)	0.54545	0.461536	0.5	13
Arrest(AR)	0.71153	0.77083	0.74	96
TRIAL	0.78703	0.72033	0.75221	118
Kidnapping(KD)	0.6	0.461536	0.52173	13
Physical abuse(PA)	0.33333	0.33333	0.33333	6
Hostage release (RE)	0.54545	0.66666	0.6	9
Shooting (SH)	0.61702	0.76315	0.68235	38
Stabbing (ST)	0.25	0.2	0.22222	5
Execution(EX)	1.0	0.5	0.66666	2
Earthquake(EQ)	0.61538	0.88888	0.72727	9
Flood(FL)	0.69230	0.75	0.72	12
Winter Storm(IR)	1.0	0.66666	0.8	3
Storm (SR)	0.77777	1.0	0.875	7
Tropical storm(TR)	1.0	0.57142	0.72727	7
Wild fire (WF)	0.5	0.33333	0.4	3
Landslides(LS)	1.0	0.5	0.66666	2
Man made disaster (MM)	0.78461	0.796875	0.79069	64
Maritime accident (MT)	0.875	1.0	0.93333	7
Explosion (XP)	0.363635	0.57142	0.44444	7
Other (NONE)	0.413047	0.39583	0.404254	48
accuracy			0.66956	690
macro avg	0.62133	0.57994	0.58426	690
weighted avg	0.66672	0.66956	0.66372	690

Table 6: Example of paraphrase generation from GPT4 (arrest category)

Source sentence	GPT4 paraphrase
Man arrested after planting fake bomb in Chicago (AP)	Individual detained for setting up counterfeit explosive in Chicago (Reuters)
	Chicago law enforcement apprehends man for hoax bomb plant (BBC)
Three arrested over injured rugby player	Trio apprehended linked to wounded rugby athlete
	Three detained in connection with harm inflicted on rugby sportsman
Suspect arrested after television appeal	Individual apprehended following TV plea
	TV appeal leads to suspect's detention

Apart from the already observed size effect (more data, better prediction), a few categories are predicted with a F1 score less than 0.90: Heavy Weapons Fire, Execution and Stabbing have the worst outcome with 0.80. Undefined is at 0.88 and all the other categories are above 0.90 which represents a new state-of-the-art by a large margin. First, if we compare the overall results to the same with fine-tuned model with just the source data, we can clearly see the benefit of data augmentation,

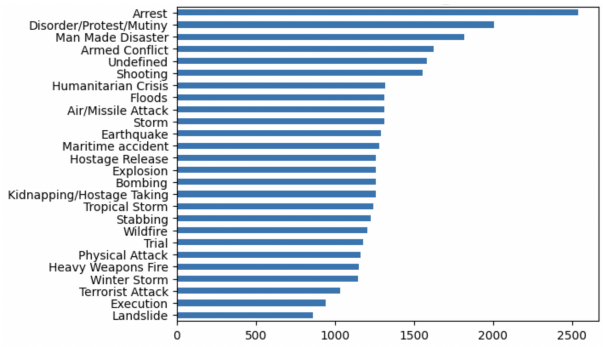


Figure 3: Distribution of classes in JRC news Dataset after GPT4 augmentation

even for the under-represented categories. (see categories with less than 100 support).

5. Conclusions and Perspectives

We have evaluated two event detection system, the first based on rules and the second, based on transformer-based classifiers. We have also experimented with data augmentation, using a state-of-the-art LLM.

Transformer-based classifiers gave overall comparable performance to the rule based system. Both systems show their own advantages: statistical classifiers achieve better classification accuracy (0.67 vs. 0.56). On the other hand, these classifiers show lower average F1 performance, mainly due to the imbalanced training set. This disadvantage was removed with data augmentation and dataset balancing, achieving much higher accuracy (0.93).

Going deeper into the details, statistical classifiers provided a much better F1 score for the classes which are frequent in the corpus, the TRIAL event class: 0.75 vs. 0.54 for NEXUS; Military operation (ARM): 0.51 vs. 0.37, and the NONE class, which is event reporting dead or injury, not belonging to any of the classes in the corpus, 0.40 vs. 0.20. The other case, where statistical classifier notably outperforms NEXUS is for the event type Storm (SR), 0.87 vs. 0.7, and Maritime accident (MT), 0.93 vs. 0.78. On the other hand, the NEXUS system has detected far better the following important event types: Terrorist attack (TA), 0.68 vs. 0.52, Kidnapping (KD), 0.72 vs. 0.52, Bombing (BO), 0.62 vs. 0.19, and Explosion (XP)

For most of the other classes we have similar performance between the two systems with the statistical biased towards more frequent classes and demonstrating much better overall accuracy and the rule based NEXUS with more balanced behaviour, showing a significantly higher macro average F1. Considering that both system approaches have different strong points, delivering a combined model will most likely deliver the most optimal results.

Another conclusion, based on the last experiments is that large language models can help build relevant datasets for fine-tuning transformer models on Event Extraction. Even if it is not possible so far to use LLMs directly for live detection, mainly due to the hardware requirements of such models and secondly due to the currently lower quality of open-sourced models, progress in these two areas should lead us in the future to directly use these models, as they show a amazing ability to learn from few examples. The next step would also be to complement sentence or passage classification with extracting the arguments of the events. For example, instead of just classifying *Two passenger trains collide in Egypt, killing 25* as a "Man made disaster", generate a structured extraction stating the specific disaster (collision), the participants (two passenger trains), the time (unspecified here but can be inferred from the source of the headline), location (Egypt) and the resulting damage (25 human deaths).

Chinatsu Aone and Mila Ramos-Santacruz. 2000. Rees: a large-scale relation and event extraction system. In *Sixth applied natural language processing conference*, pages 76–83.

Martin Atkinson, Jakub Piskorski, Hristo Tanev, and Vanni Zavarella. 2017a. On the creation of a security-related event corpus. In *Proceedings of the Events and Stories in the News Workshop*, pages 59–65, Vancouver, Canada. Association for Computational Linguistics.

Martin Atkinson, Jakub Piskorski, Hristo Tanev, and Vanni Zavarella. 2017b. On the creation of a security-related event corpus. In *Proceedings of the Events and Stories in the News Workshop*, pages 59–65.

Nancy Chinchor and Elaine Marsh. 1998. Muc-7 information extraction task definition. In *Proceeding of the seventh message understanding conference (MUC-7), Appendices*, pages 359–367.

Laura Chiticariu, Yunyao Li, and Frederick Reiss. 2013. Rule-based information extraction is dead! long live rule-based information extraction systems! In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 827–832.

Linguistic Data Consortium et al. 2005. Ace (automatic content extraction) english annotation guidelines for events version 5.4. 3. *ACE*.

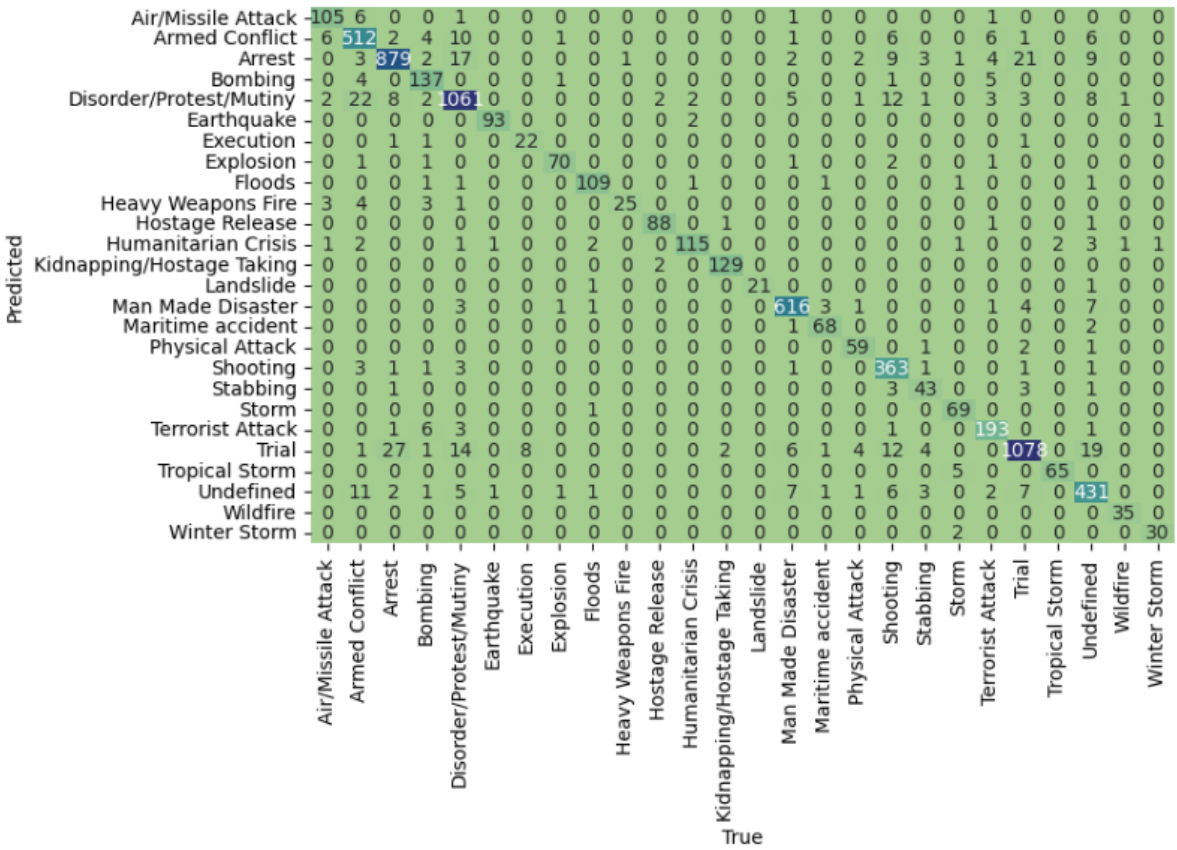


Figure 4: Dispersion plot of results of fined-tuned model on EMM dataset

Ralph Grishman, Silja Huttunen, and Roman Yangarber. 2002a. Information extraction for enhanced access to disease outbreak reports. *Journal of biomedical informatics*, 35(4):236–246.

Ralph Grishman, Silja Huttunen, and Roman Yangarber. 2002b. Real-time event extraction for infectious disease outbreaks. In *Proceedings of Human Language Technology Conference (HLT)*, pages 366–369.

Andrew Halterman, Benjamin Bagozzi, Andreas Beger, Phil Schrodt, and Grace Scraborough. 2023. Plover and polecat: A new political event ontology and dataset.

Felix Hamborg, Corinna Breitinger, and Bela Gipp. 2019. Giveme5w1h: A universal system for extracting main events from news articles. In *7th International Workshop on News Recommendation and Analytics*, pages 35–43.

Ali Hürriyetoğlu, Hristo Tanev, Vanni Zavarella, Jakub Piskorski, Reyyan Yeniterzi, and Erdem Yörük. 2021. Challenges and applications of automated extraction of socio-political events from text (case 2021): Workshop and shared task report. *arXiv preprint arXiv:2108.07865*.

Ourania Kounadi, Thomas J Lampoltshammer, Elizabeth Groff, Izabela Sitko, and Michael Leitner. 2015. Exploring twitter to analyze the public's reaction patterns to recently reported homicides in london. *PloS one*, 10(3):e0121848.

Hanspeter Kriesi, Bruno Wüest, Jasmine Lorenzini, Peter Makarov, Matthias Enggist, Klaus Rothenhäusler, Thomas Kurer, Silja Häusermann, Patrice Wangen, Argyrios Altiparmakis, et al. 2020. Poldem-protest dataset 30 european countries.

Remo Nitschke, Yuwei Wang, Chen Chen, Adarsh Pyarelal, and Rebecca Sharp. 2022. Rule based event extraction for artificial social intelligence. In *Proceedings of the First Workshop on Pattern-based Approaches to NLP in the Age of Deep Learning*, pages 71–84.

Helene Bøsei Olsen, Étienne Simon, Erik Velldal, and Lilja Øvrelid. 2024. Socio-political events of conflict unrest: A survey of available datasets. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio Political Events (CASE)*.

Jakub Piskorski, Nicolas Stefanovitch, Brian Doherty, Jens P Linge, Sopho Kharazi, Jas Mantero,

Table 7: Classification report on EMM data, fine-tuned on augmented data

	precision	recall	f1-score	support
Air/Missile Attack	0.89743	0.92105	0.90909	114
Armed Conflict	0.89982	0.92252	0.91103	555
Arrest	0.95336	0.92235	0.9376	953
Bombing	0.85625	0.92567	0.88961	148
Disorder/Protest/Mutiny	0.94732	0.93645	0.94185	1133
Earthquake	0.97894	0.96875	0.97382	96
Execution	0.73333	0.88	0.8	25
Explosion	0.94594	0.92105	0.93333	76
Floods	0.94782	0.94782	0.94782	115
Heavy Weapons Fire	0.96153	0.69444	0.80645	36
Hostage Release	0.95652	0.96703	0.96174	91
Humanitarian Crisis	0.95833	0.88461	0.92	130
Kidnapping/Hostage Taking	0.97727	0.98473	0.98098	131
Landslide	1.0	0.91304	0.95454	23
Man Made Disaster	0.96099	0.96703	0.96400	637
Maritime accident	0.91891	0.95774	0.93793	71
Physical Attack	0.86764	0.93650	0.90076	63
Shooting	0.87469	0.968	0.91898	375
Stabbing	0.76785	0.84313	0.80373	51
Storm	0.87341	0.98571	0.92617	70
Terrorist Attack	0.88940	0.94146	0.91469	205
Trial	0.96164	0.91588	0.93820	1177
Tropical Storm	0.97014	0.92857	0.94890	70
Undefined	0.87601	0.89791	0.88683	480
Wildfire	0.94594	1.0	0.97222	35
Winter Storm	0.9375	0.9375	0.9375	32
accuracy			0.93093	6892
macro avg	0.91761	0.92573	0.91991	6892
weighted avg	0.93250	0.93093	0.93113	6892

Guillaume Jacquet, Alessio Spadaro, and Giulia Teodori. 2023. Multi-label infectious disease news event corpus.

Clionadh Raleigh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. 2010. Introducing acled: an armed conflict location and event dataset: special data feature. *Journal of peace research*, 47(5):651–660.

Ralph Sundberg, Kristine Eck, and Joakim Kreutz. 2012. Introducing the ucdp non-state conflict dataset. *Journal of peace research*, 49(2):351–362.

Hristo Tanev, Jakub Piskorski, and Martin Atkinson. 2008. Real-time news event extraction for global crisis monitoring. In *Natural Language and Information Systems: 13th International Conference on Applications of Natural Language to Information Systems, NLDB 2008 London, UK, June 24-27, 2008 Proceedings 13*, pages 207–218. Springer.

Hristo Tanev, Vanni Zavarella, Jens Linge, Mijail Kabadjov, Jakub Piskorski, Martin Atkinson, and Ralf Steinberger. 2009. Exploiting machine learning techniques to build an event extraction system for portuguese and spanish. *Linguamática*, 1(2):55–66.

Michael D Ward, Andreas Beger, Josh Cutler, Matthew Dickenson, Cassy Dorff, and Ben Radford. 2013. Comparing gdelt and icews event data. *Analysis*, 21(1):267–297.

Multi-Dimensional Insights: Annotated Dataset of Stance, Sentiment, and Emotion in Facebook Comments on Tunisia's July 25 Measures

Sanaa Laabar, Wajdi Zaghouani

Hamad Bin Khalifa University
College of Humanities and Social Sciences

sala29626@hbku.edu.qa, wzaghouani@hbku.edu.qa

Abstract

On July 25, 2021, Tunisian President Kais Saied announced the suspension of parliament and dismissal of Prime Minister Hichem Mechichi, a move that sparked intense public debate. This study investigates Tunisian public opinion regarding these events by analyzing a corpus of 7,535 Facebook comments collected from the official Tunisian presidency page, specifically the post announcing the July 25 measures. A team of three annotators labeled a subset of 5,000 comments, categorizing each comment's political stance (supportive, opposing, or neutral), sentiment (positive, negative, or neutral), emotions, presence of hate speech, aggressive tone, and racism. The inter-annotator agreement, measured by Cohen's kappa, was 0.61, indicating substantial consensus. The analysis reveals that a majority of commenters supported President Saied's actions, outnumbering those who opposed or took a neutral stance. Moreover, the overall sentiment expressed in the comments was predominantly positive. This study provides valuable insights into the complex landscape of public opinion in Tunisia during a crucial moment in the country's ongoing political transformation, highlighting the role of social media as a platform for political discourse and engagement.

Keywords: Sentiment Analysis, Hate Speech, Digital Humanities, Data Annotation

1. Introduction

Eleven years have passed since Mohammed Bouazizi, a Tunisian fruit and vegetable vendor, set himself on fire in the town of Sidi Bouzid, Tunisia. This incident sparked the Arab Spring, a series of revolutions that began in Tunisia and proceeded to engulf the Arab World, including Egypt, Libya, Yemen, and Syria. In 2011, Tunisia's smooth transition into a democracy was considered a success story and a beacon of hope for democracy to its Arab neighbors who had suffered from political turmoil and civil wars after their revolutions. The Jasmine revolution erupted as a strike back to the unanswered desperate call to change the country's rising unemployment, food inflation, corruption, lack of political freedom, and poor living conditions. Demonstrations and protests spread in every governorate in the country and President Ben Ali fled through a private presidential jet to Kingdom of Saudi Arabia where he sought political refuge until his passing in 2019 in Jeddah. Unfortunately, Tunisia's democratic and progressive state ceased to exist when President Kais Said won the elections and changed the country's direction in a sudden unprecedented decision two years after his election. On the night of the 25th of July 2021, Said suspended all the works of the Tunisian parliament and dismissed his head of government, Hichem Mechichi. After living in a state of democracy for 10 consecutive years, Tunisia had suddenly been catapulted into a different political reality. This action was termed as a coup against the revolution and the constitution by his opposers; however, supporters saw it as correcting the revolution. Within one night, Tunisia witnessed a clear and sharp division amongst its people between those who expressed their support for the 25th of July measures and those who opposed them. According to Yerkes and Mbarek, "the very fact that people can freely and publicly express their criticism of the government without fear of harm or retribution is a dramatic achievement" (Yerkes & Mbarek, 2021, p.1). Therefore, the political polarity that the country witnessed literally overnight is what motivated this research and prompted the investigation to determine the dominant political stance group. This fascinating division was evident all-over social media platforms. Interestingly, among these platforms, Facebook stood out, given that it is the most popular social media platform used in the country (Statista, 2021).

The political landscape in Tunisia underwent a fundamental transformation following the revolution. After the flight of President Ben Ali, Foued Mebazaa, then Speaker of Parliament, assumed the presidential role, as detailed by Zayani (2015). Subsequently, Prime Minister Mohamed Ghannoushi signaled his intent to establish a transitional government. However, escalating protests in Al-Kasbah led to his resignation. This period marked a pivotal transition towards progressive goals, including political reform and democratic transition, as envisioned by the Committee of Safeguarding the Revolution and the Commission for Political Reform.

Expectations for a reformed Tunisia burgeoned, setting the stage for growing disenchantment with political entities, particularly in the context of the failed promises and objectives of the revolution. The emergence of Ennahda, a moderate Islamist party

Proceedings of the Second Workshop on Natural Language Processing for Political Sciences (PoliticalNLP 2024), pages 22–32

once banned, into power, and its coalition with secular parties, Congress for the Republic and the Democratic Front for Labor and Liberty, signified a significant political shift. However, the assassinations of secular opposition members Chokri Beleid and Mohamed Brahmi implicated Ennahda, exacerbating public discontent and weakening the coalition known as the Troika, amid mounting economic and social instability.

These dynamics, coupled with intensified political and ideological divisions, underscored the crumbling revolutionary objectives, as Zayani (2015) points out for over 10 years post-revolution, and led to a disillusionment with the traditional political framework, as Wolf (2019) observes. The context of perceived political failures catalyzed the public's disaffection with established political parties, influencing electoral outcomes. Mounting dissatisfaction for the political and economic status-quo grew gradually in the last 10 years prior to the 2021, 25th of July measures. Despite his lack of conventional political experience and affiliations, Kais Saied's ascension to the presidency with a significant electoral margin highlighted the populace's yearning for change. Saied's victory, as Allahoum (2019) notes, was notable given his minimal campaign efforts and outsider status in the political arena. This trust in Saied and his decisive actions can be seen as a direct response to the accumulated frustration and skepticism towards the conventional political parties and their historical legacies in Tunisia.

Consequently, as this political event is a recent event in the country's history, not enough research has been published on the topic. Hence, to study this event, a corpus that includes 5000-annotated comments retrieved from the Tunisian presidency's Facebook page through Facepager is created, specifically from the Facebook post that announced the 25th of July measures. The aim of this research is to document and analyze two groups based on their political stance toward the measures to determine Tunisians' public opinion. Moreover, our study examines the sentiment, emotion, and hate speech in the comments and their connection to a corresponding political stance. Additionally, it analyses and compares word frequencies between both groups using the tool AntConc.

2. Related Work

The current body of literature demonstrates the evolving application of digital tools in media and communication studies, as evidenced by research from Mahmadi et al. (2017) and Saad & Sabrini-Zin (2022), which illustrate the application of these tools in extracting and analyzing digital content. The evolving landscape of news dissemination underscores the significant role of digital tools in adapting to the changing patterns of news consumption. Within this context, the United States elections emerge as a focal point, attracting extensive coverage and prioritization in news monitoring efforts. Mhamdi et al. (2017) illustrated this by employing Facepager to gather data from the Facebook pages of prominent news channels, namely Fox News, CNN, and ABC News. This data was subsequently analyzed using the same tool to understand the dynamics of digital news dissemination. Additionally, the analytical capabilities of digital tools extend beyond news analysis, as demonstrated by Saad & Sabrini-Zin (2022), who utilized AntConc to examine the lexical features of Robert Frost's poem \"Into My Own.\" These instances highlight the multifaceted applications of digital tools in both news analysis and literary studies, reflecting their integral role in contemporary research methodologies. The study demonstrated that the use of the tool facilitated enhanced insights compared to earlier research reliant solely on manual analysis. Despite the qualitative nature of the study's methodology, a descriptive approach was employed to quantitatively analyze the data extracted from the text. This methodological combination allowed for a more nuanced understanding of the textual information, showcasing the advantage of integrating digital tools in the analytical process.

To assess public sentiment regarding the July 25 measures via traditional news media, multiple surveys were executed. Shems FM, a local Tunisian News Agency, surveyed 1,707 individuals aged 18 and above from August 4 to 13. The findings showed substantial support: 87.3% of the respondents backed the decisions made on that date, 81.6% endorsed the suspension of the parliament, and 76.2% agreed with the dismissal of Prime Minister Hichem Mechichi. In a separate study by L'Economiste Maghrebin, 94.9% of participants expressed support for these exceptional measures (Marzouk, 2021). Additionally, Sigma Conseil conducted a poll reflecting a 72.2% approval rate for President Kais Saied and his policies, further indicating significant public endorsement of the actions taken on and following July 25 (2021).

Our methodology is informed by Zaghouani and Awad's (2016a) work on developing an Arabic punctuated corpus, which provides critical insights into annotation guidelines. Similarly, the comprehensive framework for annotating Arabic corpora for machine translation, as elucidated by Zaghouani et al. (2016), offers a robust model for our annotation processes. Our dataset creation approach draws upon Hawwari et al.'s (2016) meticulous annotation of Arabic morphological patterns, presenting a nuanced understanding of linguistic intricacies. Zaghouani et al.'s (2010) work on the revised Arabic Propbank further guides our dataset structuring, emphasizing the importance of detailed proposition annotation in political discourse analysis.

Bianchi et al. (2023) explore digital communication's nuances, shedding light on interactive dynamics relevant to our analysis of social media political framing. Biswas et al.'s (2023) examination of Twitter content for vaccine-related discussions exemplifies the potential of social media analytics in extracting meaningful insights from online discourse, paralleling our sentiment and sarcasm detection efforts as discussed by Farha et al. (2021).

Moreover, the foundational work by Obeid et al. (2016) on Arabic diacritization and the structured annotation processes outlined by Zaghouani and Awad (2016b) significantly influence our annotation guideline development. The creation of a multi-genre Arabic corpus by Bouamor et al. (2016) and the dataset focusing on political framing in the U.S. COVID-19 discourse by Shurafa et al. (2020) provide methodological blueprints for our data compilation and analysis efforts.

Our research is further contextualized within the broader discourse of social media data analysis. The extraction and examination of dialectal Arabic irony from Twitter, as conducted by Abbes et al. (2020), and the detection of propaganda in Arabic content, as explored by Alam et al. (2022), offer pertinent insights into the complexities of online political communication. The multi-dialect Twitter corpus analysis by Zaghouani and Charfi (2018a) enriches our understanding of language variety and demographic factors in social media interactions.

In synthesizing these diverse studies, our research aims to extend the existing scholarly discourse on political framing and annotation within the dynamic realm of social media, leveraging the rich corpus of Arabic language content and the multifaceted methodologies established in the aforementioned works.

3. Methodology

In this investigation, the Facepager application, conceived in 2019 by communication scientists Jakub Jünger of the University of Greifswald and Till Kelling of Ludwig-Maximilians University, served as the primary tool for data acquisition. Specifically, the study targeted Facebook comments extracted from a post on the official Tunisian presidency's Facebook page. This post pertained to the unprecedented suspension of parliament, a pivotal event in Tunisian political discourse. The official Facebook page of the Tunisian presidency, established in 2011 and becoming active in content posting since 2012 under President Moncef Marzouki, acts as a crucial digital platform for presidential communication. Its significance has been consistent through the administrations of Presidents Moncef Marzouki, Beji Caid Essebsi, and the incumbent president, Kais Said, marking it as a significant source of official presidential communications and public engagement. For the data extraction process, a MacBook Air running the macOS Big Sur operating system was utilized to operate the Facepager software.

This tool facilitated the efficient extraction of relevant comment data, which is typically exported into a .csv format suitable for analysis in spreadsheet applications like Macintosh's Numbers. However, to streamline the analytical workflow and enhance the ease of data annotation, the extracted Facebook comments were systematically transferred to a Google spreadsheet. This transition allowed for a more manageable and interactive engagement with the data.

Complementing the data collection process, the study incorporated the digital text analysis tool AntConc. This software, developed by Dr. Laurence Anthony, a professor and software engineer at Waseda University, Japan, is instrumental in performing a comprehensive analysis of textual data. AntConc's sophisticated functionalities support a deep dive into the linguistic and thematic elements of the extracted comments, facilitating a nuanced exploration of public sentiment and discourse patterns. By integrating these methodological tools, the research aims to meticulously parse the digital public sphere's reactions and interactions concerning the significant political development of parliament suspension in Tunisia, thus enabling a rich and insightful examination of the public discourse captured on this official digital platform.

3.1 Linguistic Variations in Tunisian Social Media Discourse

The Tunisian social media landscape is characterized by unique linguistic variations, particularly the use of "Arabizi" and code-switching between Arabic, French, and English. Arabizi refers to the romanization of Arabic text, where users employ Latin script to write Arabic words and phrases (Bjørnsson, 2010). This phenomenon is prevalent in informal online communication, especially among younger generations. Additionally, code-switching, the alternation between multiple languages within a single conversation or utterance, is a common practice in Tunisian social media discourse (Kebeya, 2013). These linguistic variations pose challenges for sentiment analysis, as they introduce non-standard orthography and complicate the identification of language-specific features. In our corpus, we observed several instances of Arabizi and code-switching, such as "sbe7 el khir" (good morning) and "vive la Tunisie" (long live Tunisia). Future research could explore methods to effectively handle these linguistic phenomena in sentiment analysis tasks.

3.2 Data Collection and Sample Representativeness

The Facebook comments analyzed in this study were collected on [insert date] from the official Tunisian Presidency Facebook page, specifically focusing on the post announcing the July 25 measures. A total of 7,535 comments were retrieved, out of which 5,000 were randomly selected for annotation. While the random sampling method aimed to ensure an unbiased representation of the overall comment population, it is important to acknowledge that Facebook users may not be entirely representative of the Tunisian general public. According to the World Bank (2021), around 66% of the Tunisian population had access to the internet in 2020, suggesting that a significant portion of the population may not be active on social media platforms. Additionally, the demographic composition of Facebook users in Tunisia may skew towards younger age groups and those with higher digital literacy. Future research could benefit from incorporating demographic data of the commenters, if available, to assess the

representativeness of the sample and potential biases in the data.

3.3 Clarification on Stance and Emotion Annotation

In our study, stance and emotion were annotated independently by the human annotators. Stance was categorized as either pro-Saied (supporting the July 25 measures), anti-Saied (opposing the measures), or neutral. Emotions, on the other hand, were annotated based on the presence of specific emotional cues in the comment text, such as joy, trust, anger, or fear. The phrase "stance count by predominant emotions" in the Results section refers to the distribution of stance categories (pro, anti, or neutral) within each emotion category. For example, among the comments annotated with the "joy" emotion, we examined the proportion of pro-Saied, anti-Saied, and neutral stances. This analysis aimed to uncover potential correlations between specific emotions and political stances. However, it is crucial to note that emotions were not used as determinants of stance; rather, they were treated as separate but potentially related dimensions of the public opinion landscape.

3.4 Data Collection

The data for this study were collected using the Facepager tool, specifically targeting comments from the official announcement regarding the temporary 30-day suspension of Tunisia's Assembly of the Representatives of the People. As depicted in Figure 1, this announcement was made on the official presidency page, attracting significant interaction, evidenced by the retrieval of 7,572 comments. Due to time constraints, a subset of 5,000 comments was selected for annotation.

Figure1: Presidency Facebook Page

Moreover, data collection utilized AntConc, a tool that necessitates a reference corpus to activate features such as "Word List" and "Key Word List." For this purpose, a corpus comprising 400 articles from Economic Arabic Newspapers (Al-Sluaiti & Atwell, 2003) was employed. Upon processing this corpus, AntConc generated a keyword list highlighting terms with unexpectedly high frequency compared to those in the reference corpus and identified collocates to reveal words frequently associated with the search terms, assessing the association's strength.

Additionally, manual annotation was conducted to differentiate between subjective sentiments and objective facts, incorporating thematic analysis to evaluate emotions and feelings with attention to subjectivity and complexity (Monkey Learn, 2022; Williams et al., 2019). Hate speech detection adhered to the criteria from Hate.org, focusing on malicious intentions towards specific groups (Crabb et al., 2019). Despite the inherent subjectivity in such analysis, the pre-trained annotators achieved a kappa score of 0.61, signifying substantial inter-annotator agreement. This was based on a blind sample of 200 comments, indicating a high consensus among the annotators in determining sentiment and emotion.

4. Results

The findings from this study provide a nuanced understanding of public sentiment, emotions, and stances expressed in the Tunisian political discourse, particularly following the July 25 measures.

4.1 Sentiment Analysis

This section illustrates a quantification of sentiments extracted from the Facebook comments. According to Google sheets findings, the sentiment results in Figure 2 show a higher value in "Very Positive" by 2,711 comments and "Positive" by 1,180 comments. On the other hand, the "Very Negative" sentiment is expressed in 307 comments, and the "Negative" sentiment in 112 comments, which are significantly less in comparison. The sentiment analysis reveals a predominantly positive public opinion towards the July 25 measures, with the combined "Very Positive" and "Positive" sentiments accounting for a substantial majority of the analyzed comments. This finding suggests a broad base of support for President Saied's actions among the Facebook users who engaged with the official presidency page.

4.2 Emotions and Feelings Analysis

The analysis of feelings and emotions, as shown in Figure 3, reveals that "Trust" and "Joy" are the predominant emotions, reflecting a generally favorable public sentiment towards the political developments. The value of "Trust" is expressed in 2,883 comments, whereas "Joy" is present in 832 comments. "Neutral" emotions are found in 597 comments, and "Anger" is detected in only 353 comments. The prevalence of "Trust" and "Joy" in the analyzed comments underscores the public's confidence in and enthusiasm for the measures taken by President Saied. These positive emotions align with the overall supportive stance towards the July 25

events, as evidenced by the sentiment analysis results.

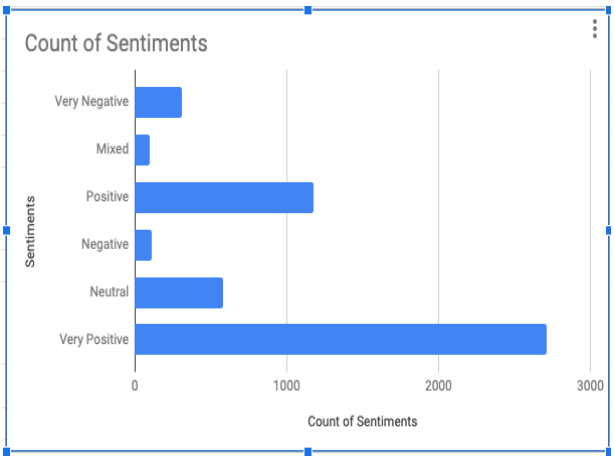


Figure 2: Count of Sentiments

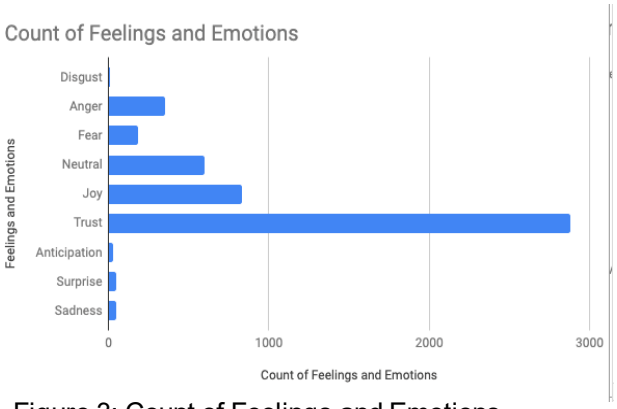


Figure 3: Count of Feelings and Emotions

4.3 Stance Analysis

The predominant emotions are reflected in the stance count, as shown in Figure 4, where a significant majority of comments supported the July 25 measures, with neutral and opposing stances being less prevalent.

The count of stance indicates that most comments, 78.6% (3,922 comments), prove to be "pro" the 25th of July measures. Moreover, the "Neutral" stance, at 12.9% (646 comments), is more common than the "Against" stance, which accounts for only 8.4% (422 comments).

The stance analysis confirms the overwhelming public support for President Saied's actions, with the "pro" stance dominating the discourse on the official presidency Facebook page.

The relatively low proportion of "Against" comments suggests that the opposition to the measures was limited, at least within the scope of this study's dataset.

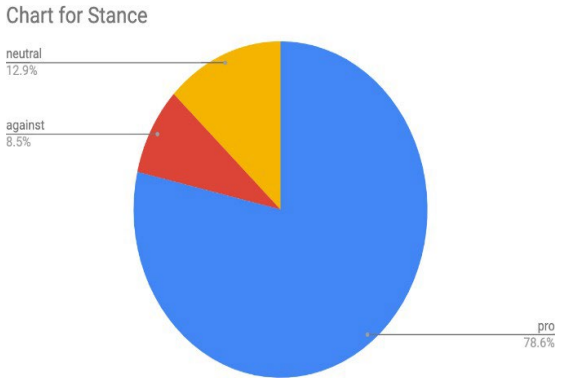


Figure 4: Count of Stance

4.4 Hate Speech, Aggressive Tone, and Racism Analysis

Despite the overall supportive sentiment, annotators assessed the presence of hate speech, aggressive tone, and racism in each comment, assigning a binary value of 'yes' or 'no'. Figure 5 shows the analysis, which revealed that a significant majority, 83.2% (4,076 comments), did not exhibit these negative tones, while only 16.6% (924 comments) did.

Notably, among the comments identified with hate speech, aggressive tone, or racism, the majority were supportive ('Pro') in stance, totaling 835 comments. Conversely, 79 comments were opposed ('Against'), and a mere 10 comments were categorized as 'Neutral'. This distribution underscores the predominance of such negative tones in supportive comments within the dataset.

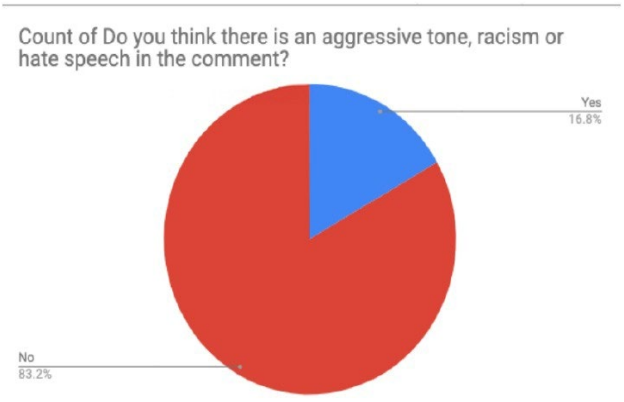


Figure 5: Count of Hate Speech, Aggressive Tone, and Racism

The analysis of hate speech, aggressive tone, and racism provides a more nuanced perspective on the public discourse surrounding the July 25 measures. While the overall sentiment was largely positive and supportive, the presence of these negative elements, particularly within the "pro" stance comments, highlights the potential for polarization and the expression of extreme views, even among supporters of President Saied's actions.

4.5 AntConc Analysis

The complexity of the public discourse is further explored in AntConc's word list analysis. By examining the corpus using AntConc, the findings show that the word "Tunis تونس" occurred 1,836 times, "people 1,234 "الشعب times, "President 1,106 "الرئيس times, and "with you 826 "معاك times. However, in the keyword list, it is noticeable that the word "Coup D'etat انقلاب" only occurred 284 times, and the word "army 191 "الجيش times.

To observe the differences in words associated with President Kais Saied and the most popular Islamic political party in the country, the researcher input common words from the corpus, such as "سعيد" ("Said," the last name of President Kais Saied), "Ennahdha النهضة," and "Brotherhood خوانجية." The collocates linked to President Saied convey positive sentiments with phrases like "elevate you يرفعك," "protect him يحفظو," and "we support نساند," indicating support. Conversely, collocates related to the Ennahdha party "النهضة" and the Brotherhood "خوانجية" are negative, showing aggression and hate speech with terms like "they destroy it يهلكوها," "they kill us يقتلونا," and "arrest them يعتقلوهم." Negative words such as "terrorists إرهابيين," "they stole it سرقوه," and "they destroyed it دمروه" are also associated with these groups, highlighting a stark contrast in sentiment.

4.6 Keyword and Collocates Analysis

The results of the keyword list analysis, presented in Table 1, display an interesting set of key terms that were predominant in the corpus. Keywords that apparently favor Saied and his decisions rank highest in keyness and frequency, while the word "Coup" has the lowest value.

Furthermore, the collocates match the percentage of stance and carry positive sentiments, emotions of joy, and evident support towards President Kais Saied, as expected. On the other hand, the term "Coup" seems to carry a majority of collocates that have neutral stances or sentiments, with the remaining collocates expressing positive sentiments that indicate a supportive stance towards President Saied.

4.7 Corpus and Annotation Challenges

Finally, in contextualizing the corpus size and annotation challenges, it should be noted that the corpus comprises 7,535 Facebook comments, of which 5,000 were annotated across five dimensions: comment text, sentiment, stance, emotions, and hate speech. The annotations were organized in two Excel tabs, providing a detailed view of the average pairwise annotation agreements between the three annotators for sentiment and emotion categories. A recurrent issue in the annotation process was the determination of sentiment and emotion intensity levels.

Divergences in annotation were particularly notable in the assessment of sentiments and emotions. Annotators often disagreed on whether a comment should be classified as negative or very negative and in distinguishing between emotions like trust and joy. These discrepancies were largely due to differences in how annotators perceived the commenter's intentions and emotional responses to President Saied's actions. For example, the comment "# لا للانقلاب No to Coup" was deemed negative by one researcher, who argued that the comment lacked the linguistic intensity to be categorized as very negative. Conversely, a second annotator interpreted the same comment as very negative, illustrating the subjective nature of interpreting sentiment and emotion in textual analysis.

Rank	Freq	Keyness	Effect	Keyword
1	2652	+ 4592.17	0.069	و (and)
2	1836	+ 4275.02	0.0485	تونس (Tunis)
3	1234	+ 2856.08	0.0328	الشعب (The people)
4	1106	+ 2143.7	0.0294	الرئيس (The president)
5	826	+ 1949.72	0.0221	معاك (With you)
6	814	+ 1906.63	0.0218	يا (O')
7	699	+ 1649.11	0.0187	تحيا (Long live)
8	606	+ 1429.17	0.0163	شكرا (Thank you)
9	611	+ 1415.56	0.0164	قيس (Kais)
10	469	+ 1105.47	0.0126	ربي (My God)
11	629	+ 1046.08	0.0169	الله (Allah)
12	415	+ 977.98	0.0112	سيدي (Mr.)
13	325	+ 765.61	0.0088	سيادة (His excellency)
14	536	+ 746.46	0.0144	سعيد (Said)
15	284	+ 668.92	0.0077	انقلاب (Coup D'etat)

Table 1: Keywords List

5. Discussion

The findings from this study offer valuable insights into the complex landscape of public opinion and sentiment in the aftermath of the July 25 measures taken by President Kais Saied in Tunisia. The analysis of Facebook comments from the official Tunisian presidency page reveals a multifaceted

discourse, with a predominant sentiment of support for the president's actions.

5.1 Sentiment Analysis

The sentiment analysis results clearly demonstrate that the majority of Tunisians who commented on the July 25 measures, specifically the suspension of parliament ordered by President Kais Saied, were in support of his decision. The stance analysis shows that the pro-decision group constitutes the majority, while those against the measures form a minority, even smaller than the neutral group. This finding provides readers with a clear understanding that the opposition to the measures was limited within the analyzed dataset. Furthermore, the sentiment analysis reveals a dominant score in the "Very Positive" category, followed by "Positive" and "Neutral," with "Negative" and "Very Negative" sentiments scoring significantly lower. This observation leads to the conclusion that the general sentiment surrounding the July 25 measures is highly positive, with those expressing negative sentiments being considered a minority.

5.2 Emotions and Feelings Analysis

The emotions and feelings analysis aligns with the sentiment findings, indicating that many Tunisians placed high trust in Saied's decision, as evidenced by the large number of comments scoring in the "Trust" value. The prominence of the "Joy" value also suggests that a significant portion of Tunisians were genuinely happy about these sudden decisions, which may explain the large crowds that took to the streets in the middle of the night to celebrate. Conversely, the values of "Anger" and "Fear" are far less prevalent in the analyzed comments, implying that individuals harboring these feelings were fewer in number. These latter values would primarily be associated with those who opposed the decision, further reinforcing the correlation between the emotions count and the stance count results.

5.3 Hate Speech, Aggressive Tone, and Racism Analysis

The analysis of hate speech, aggressive tone, and racism yields an intriguing finding. While the overall incidence of these negative expressions is lower, it is noteworthy that they are more frequent in comments from the pro-decision faction. Despite the corpus showing minimal hate speech or racism, with most comments expressing joy and trust towards the president's decisions, the pro-decision group, despite being the majority, displayed more hate speech than those against the measures. This observation suggests that alongside the positive feelings, the pro-decision group also harbored resentment towards other political entities.

5.4 AntConc Analysis

The AntConc findings further corroborate the sentiment and stance results, with the lexical analysis revealing a predominance of words conveying positive support for the president and the nation. The frequent occurrence of terms such as "Tunis," "President," "Saied," and "People" underscores this sentiment. The analysis of collocates introduces a nuanced dichotomy reflective of Tunisia's past and present political landscape. Collocates associated with the term "Saied" carry positive connotations, indicating broad support and approval of President Saied's measures.

In contrast, the analysis of "Ennahdha" and "Brotherhood" reveals a marked hostility towards these entities. The term "خوانجية" (Brotherhood), often used derogatorily in Tunisian political discourse, frequently appears in collocates with negative sentiment. This term is unfavorably associated with the Ennahdha party and its affiliates, who typically reject this nomenclature. The name "Ennahdha" (نهضة), representing the country's most prominent Islamic political party, is widely recognized and used in a non-derogatory manner by the general population. The linguistic evidence suggests a pervasive disdain for this party, extending beyond the events of July 25 and reflecting longstanding political tensions. Ennahdha's significant role in Tunisia's political arena, coupled with its visibility and cohesive group identity, has evidently fueled the negative sentiment captured in the study's corpus. This trend of animosity, primarily directed at Ennahdha and its representatives, indicates a polarized political sentiment within the Tunisian populace.

5.5 Keyword and Collocates Analysis

The exploration of keywords and collocates reinforces the narrative of widespread support for President Saied and the measures taken on July 25. Keywords such as "president" (الرئيس), "his excellency" (سيادة), "the people" (الشعب), and "Tunis" highlight a positive sentiment towards the July 25 measures, focusing on support for Tunisia and President Saied. Positive collocates associated with "The President" suggest widespread approval. Conversely, the term "Coup," used by the opposition, ranks low in the keyword list, indicating that those against the measures are a minority. While some collocates of "Coup" show negative sentiment, others are neutral to positive, suggesting varied perceptions even among supporters, who use the term either critically or in support.

5.6 Implications and Future Directions

The findings of this study have significant implications for understanding the complex dynamics of public opinion in the context of Tunisia's ongoing political transformation. The overwhelming support for President Saied's measures, as expressed in the analyzed Facebook comments, suggests a strong public mandate for his actions and a desire for change in the country's political landscape. However, the presence of hate speech and aggressive tone, particularly among the pro-decision group, highlights the potential for polarization and the need for fostering a more inclusive and respectful public discourse.

Future research could explore the evolution of public sentiment over time, as the political situation in Tunisia continues to unfold. Longitudinal studies could provide valuable insights into how opinions and

emotions shift in response to specific events and policy decisions. Additionally, the incorporation of demographic data, if available, could help identify any differences in sentiment and stance across various segments of the Tunisian population.

Moreover, the linguistic analysis could be expanded to include a more comprehensive examination of the Tunisian dialect and its unique features, such as the use of Arabizi and code-switching. This would enable a more nuanced understanding of the language used in online political discourse and its potential impact on sentiment analysis and opinion mining.

6. Limitations

This study faced two primary limitations: temporal constraints and corpus size. The restricted timeframe necessitated limiting the sample to 5,000 annotated Facebook comments, constraining the breadth of analysis and precluding broad generalizations. Consequently, this research should be regarded as an initial, exploratory, small-scale study. Despite these limitations, the pilot nature of this work lays the groundwork for future, more expansive research endeavors. Importantly, the manual annotation performed in this study serves as a valuable precursor to the development of machine learning tools and algorithms tailored for detecting sentiment polarity and stance within Tunisian dialect corpora, thereby enhancing the methodological approach to analyzing this linguistic context. Moreover, due to the limitation of time and corpus size, the researcher's annotated notes only included 35 comments that were written in 'Arabizi', the Arabic chat alphabet, of the Tunisian dialect. Evidently, the number of comments was ostensibly small and therefore the researchers we not able to include it in the linguistic part of the data analysis. However, the 'Arabizi' comments were included in the stance and emotion data collection and analysis.

7. Conclusion

The research indicated a prevalent support among Tunisian Facebook users for the July 25 measures, with "Trust" and "Joy" being the predominant emotions expressed. Opposition to these measures was comparatively minor, with neutral stances more common than outright negative ones. Notably, the "pro" faction, while largely supportive, exhibited a greater tendency towards hate speech and aggressive tones, though these instances were relatively few. The prevailing sentiment among the comments was "Very Positive," reflecting a broad endorsement of the actions taken.

Linguistic analysis, including word lists and collocates, pointed to significant backing for President Saied, alongside notable criticism of the "Ennahdha" party, underscoring a clear political divide. This study highlights the necessity for more nuanced inquiries into the shifting sentiments of Tunisians regarding President Saied's policies, particularly through the lens of symbolic and emotive elements like emojis. The July 25 measures emerge as a crucial juncture in Tunisia's political and democratic trajectory, meriting deeper examination of the public's reaction and its implications for the country's future.

The findings of this study contribute to the growing body of research on public opinion and sentiment analysis in the context of political events, particularly in the Middle East and North Africa region. The study's focus on Tunisia's July 25 measures provides valuable insights into the complex dynamics of public discourse and the role of social media in shaping political narratives. The prevalence of support for President Saied's actions, as evidenced by the analyzed Facebook comments, underscores the importance of understanding public sentiment in times of political upheaval.

However, the study also reveals the potential for polarization and the presence of hate speech and aggressive tones, even among supporters of the measures. This finding highlights the need for further research into the factors contributing to the spread of such negative sentiments and the development of strategies to promote a more inclusive and respectful public discourse.

8. Data Availability Statement

The annotated dataset can be obtained by contacting the authors to facilitate future research and reproducibility. The users of the dataset must adhere to the terms and conditions outlined in the repository. To request the dataset for research purposes, please fill the following form:

https://forms.gle/S9fZtYjAyLAqFsH19

The data are not publicly available due to the sensitive nature of the comments and the potential for misuse or misinterpretation outside the context of this research. Access to the data will be granted to researchers. Requestors will be required to sign a data sharing agreement that specifies the conditions under which the data can be used, including measures to protect the privacy and confidentiality of the individuals whose comments are included in the dataset. The annotated dataset will be made available in a de-identified format, with any personally identifiable information removed to ensure the anonymity of the commenters.

Researchers interested in replicating or building upon the findings of this study are encouraged to contact the authors to discuss data access and collaboration opportunities.

9. Ethical Statement

This study was conducted in accordance with the ethical guidelines and regulations set forth by Hamad Bin Khalifa University. The study involved the analysis

of publicly available data from the official Facebook page of the Tunisian presidency.

As such, the research did not involve any direct interaction with human subjects and did not require informed consent from the individuals whose comments were included in the dataset. However, the researchers recognize the potential for harm and the need to protect the privacy and confidentiality of the commenters, even in the context of publicly available data.

- To mitigate potential risks and ensure the ethical conduct of the research, the following measures were taken:
- The data were collected and analyzed in an anonymous and de-identified format, with any personally identifiable information removed from the dataset.
- The researchers have taken steps to secure the data and prevent unauthorized access, including storing the data on encrypted drives and limiting access to authorized personnel only.
- The findings of the study are reported in aggregate form, without singling out or identifying any individual commenters.
- The researchers have strived to present the findings in a balanced and objective manner, avoiding any stigmatization or stereotyping of individuals or groups based on their opinions or political affiliations.
- The researchers are committed to the ethical and responsible conduct of research and have taken these measures to ensure that the study complies with the highest standards of academic integrity and human subjects protection. Any concerns or questions about the ethical aspects of this study should be directed to the authors.

10. Acknowledgments

This publication was made possible by NPRP14C-0916-210015 / MARSAD Sub-Project from the Qatar National Research Fund / Qatar Research Development and Innovation Council (QRDI). The contents herein reflect the work and are solely the authors' responsibility.

11. References

Abbes, I., Zaghouani, W., El-Hardlo, O., & Ashour, F. (2020). DAICT: A dialectal Arabic irony corpus extracted from Twitter.

Alam, F., Mubarak, H., Zaghouani, W., Martino, G. D. S., & Nakov, P. (2022). Overview of the WANLP 2022 shared task on propaganda detection in Arabic.

Al-Ghadir, A., Azmi, A., & Hussain, A. (2021). A novel approach to stance detection in social media tweets by fusing ranked lists and sentiments. *Information Fusion, 67*, 29-40. https://doi.org/10.1016/j.inffus.2020.10.003

Al-Sulaiti, L., & Atwell, E. S. (2003). *The Design of a Corpus of Contemporary Arabic (CCA)*. The University of Leeds.

Aldayel, A., & Magdy, W. (2019). Your stance is exposed! Analyzing possible factors for stance detection on social media. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW), Article 1. https://doi.org/10.1145/3359307

Allahoum, R. (2019, October 11). Tunisia's Kais Saied: 'He's just not interested in power'. *Aljazeera*. https://www.aljazeera.com/news/2019/10/11/tunisias-kais-saied-hes-just-not-interested-in-power

Baraniak, K., & Sydow, M. (2021). A dataset for sentiment analysis of entities in news headlines (SEN). *Procedia Computer Science, 192*, 3627-3636. https://doi.org/10.1016/j.procs.2021.09.136

Bianchi, R., Weber, A., Yyelland, B., Ghanam, R., Kittaneh, K., & Zaghouani, W. (2023). SYNCHRONOUS CONFERENCING SOFTWARE-ASSISTED TEACHING & LEARNING.

Biswas, M. R., Mohsen, F., Shah, Z., & Zaghouani, W. (2023). Potentials of ChatGPT for Annotating Vaccine Related Tweets.

Bothe, C., Weber, C., Magg, S., & Wermter, S. (2019). EDA: Enriching emotional dialogue acts using an ensemble of neural annotators. *arXiv preprint arXiv:1912.00819*.

Bouamor, H., Zaghouani, W., Diab, M., Obeid, O., Oflazer, K., Ghoneim, M., & Hawwari, A. (2016). On Arabic Multi-Genre Corpus Diacritization.

Crabb, J., Yang, S., & Zubova, A. (2019, December 10). Classifying hate speech: An overview. *Medium*. https://towardsdatascience.com/classifying-hate-speech-an-overview-d307356b9eba

Espace Manager. (2021) Sigma conseil: Kaïs Saïed caracole loin en tête avec 72,2% de confiance et 90% d'intentions de vote. *Espace Manager.*

https://www.espacemanager.com/si gma-conseil-kais-saied-caracole- loin-en-tete-avec-722-de-confiance- et-90-dintentions-de-vote.html.

Farha, I. A., Zaghouani, W., & Magdy, W. (2021). Overview of the WANLP 2021 shared task on sarcasm and sentiment detection in Arabic.

Gambino, O., Calvo, H., & García-Mendoza, C. (2018). Distribution of emotional reactions to news articles in Twitter. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*

(LREC 2018). https://doi.org/979-10-95546-00-9

Hardalov, M., Arora, A., Nakov, P., & Augenstein, I. (2021). A survey on stance detection for mis- and disinformation identification. *arXiv preprint arXiv:2103.00242*.

Hasan Saad, A., & Sarbini-Zin, M. L. (2022). Corpus-assisted analysis of Robert Frost's poem, "Into My Own" using AntConc. *PENDETA, 13*(1), 1-9.

Hawwari, A., Zaghouani, W., Diab, M., O Gorman, T., & Badran, A. (2016). Ampn: a semantic resource for Arabic morphological patterns.

Hendrickson, C., & Galston, W. (2017, April 28). Why are populists winning online? Social media reinforces their anti-establishment message. *Brookings*. https://www.brookings.edu/blog/techtank/2017/04/28/why-are-populists-winning-online-social-media-reinforces-their-anti-establishment-message/

Jungherr, A., Posegga, O., & An, J. (2021). Populist supporters on Reddit: A comparison of content and behavioral patterns within publics of supporters of Donald Trump and Hillary Clinton. *Social Science Computer Review, 40*, 089443932199613. https://doi.org/10.1177/0894439321996130089443932199613.

Jurafsky, D. (1997). *Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13* (Technical Report 97-01, pp. 225–233). University of Colorado Institute of Cognitive Science.

Aljazeera. (2019). Kais Saied wins Tunisia election with 72 percent. https://www.aljazeera.com/news/2019/10/14/tunisia-presidential-election-kais-saied-declared-winner

Tunisian Presidency. (2022). Tunisian presidency Facebook comments 25th of July measures. [Publisher unknown]. URL

Lai, M., Cignarella, A., Hernández Farías, D., Bosco, C., Patti, V., & Rosso, P. (2020). Multilingual stance detection in social media political debates. *Computer Speech & Language, 63*, 101075. https://doi.org/10.1016/j.csl.2020.101075

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLoS ONE, 14*(8), e0221152. https://doi.org/10.1371/journal.pone.0221152

Marzouk, H. (2021, August 17). Baromètre politique aout 2021 Tunisiens soutiennent mesures exceptionnelles 25 Juillet. *L'Economiste Maghrébin*. https://www.leconomistemaghrebin.com/2021/08/17/barometre-politique-aout-2021-949-tunisiens-soutiennent-mesures-exceptionnelles-25-juillet/

Mhamdi, C., Al-Emran, M., & Salloum, S. A. (2018). Text mining and analytics: A case study from news channels posts on Facebook. In *Intelligent Natural Language Processing: Trends and Applications* (pp. 399-415). Springer.

Mutlu, E., Oghaz, T., Jasser, J., Tutunculer, E., Rajabi, A., Tayebi, A., Ozmen, O., & Garibay, I. (2020). A stance data set on polarized conversations on Twitter about the efficacy of hydroxychloroquine as a treatment for COVID-19. *Data in Brief, 33*, 1-11. https://doi.org/10.1016/j.dib.2020.106401

Obeid, O., Bouamor, H., Zaghouani, W., Ghoneim, M., Hawwari, A., Alqahtani, S., Diab, M., & Oflazer, K. (2016). Mandiac: A web-based annotation system for manual Arabic diacritization.

Saif, H., & Alani, H. (2012). Semantic sentiment analysis of Twitter. In *The Semantic Web--ISWC 2012* (Vol. 7649, pp. 508-524). https://doi.org/10.1007/978-3-642-35176-1_32

Sandoval-Almazan, R., & Valle-Cruz, D. (2018). Facebook impact and sentiment analysis on political campaigns. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age* (pp. 1-7)

Shems FM. (n.d.). Sondage: 87,3% des Tunisiens approuvent les décisions de Kais Saied, annoncées le 25 juillet. https://www.shemsfm.net/amp/fr/actualites_tunisie-news_news-nationales/309135/sondage-87-3-des-tunisiens-approuvent-les-decisions-de-kais-saied-annoncees-le-25-juillet.

Shurafa, C., Darwish, K., & Zaghouani, W. (2020). Political framing: US COVID19 blame game.

Statista. (2022). Tunisia: Share of social media users by platform. https://www.statista.com/statistics/1196465/share-of-active-social-media-users-by-platform-tunisia/.

Vashisht, G., & Thakur, S. (2014). Facebook as a corpus for emoticons-based sentiment analysis. *International Journal of Emerging Technology and Advanced Engineering, 4*, 904-908.

Wang, J., & Chen, K. (2021). Aggregating user-centric and post-centric sentiments from social media for topical stance prediction. *ROCLING*.

Williams, L., Arribas-Ayllon, M., Artemiou, A., et al. (2019)

Comparing the utility of different classification schemes for emotive language analysis. *Journal of Classification, 36*, 619–648. https://doi.org/10.1007/s00357-019-9307-0

Wolf, A. (2019). In search of 'consensus': The crisis of party politics in Tunisia. *The Journal of North African Studies, 24*(6), 883-886. https://doi.org/10.1080/13629387.2019.1675249

Zamani, N. A. M., Abidin, S. Z., Omar, N. A. S. I. R. O. H., & Abiden, M. Z. Z. (2013). Sentiment analysis: Determining people's emotions in Facebook. In *Proceedings of the 13th International Conference on Applied Computer and Applied Computational Science* (pp. 111- 116).

Zayani, M., & Downing, J. (2015). *Networked Publics and Digital Contention* (1st ed.). Georgetown University's Center of International and Regional Studies, School of Foreign Services in Qatar.

Shems FM. (2021). Sondage: 87,3% des tunisiens approuvent les décisions de Kais Saied, annoncées le 25 juillet. https://www.shemsfm.net/fr

Pak, A., & Paroubek, P. (2021, May). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

Baj-Rogowska, A. (2017, December). Sentiment analysis of Facebook posts: The Uber case. In *2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS)* (pp. 391-395). IEEE.

Abercrombie, G., & Batista-Navarro, R. T. (2020, May). ParlVote: A corpus for sentiment analysis of political debates. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 5073-5078).

Yerkes, S., & Mbarek, N. (2021, January 14). After ten years of progress, how far has Tunisia really come? Carnegie Endowment for International Peace. https://carnegieendowment.org/2021/01/14/after-ten-years-of-progress-how-far-has-tunisia-really-come-pub-83609

Zaghouani, W., & Awad, D. (2016a). Building an Arabic Punctuated Corpus.

Zaghouani, W., & Awad, D. (2016b). Toward an Arabic punctuated corpus: Annotation guidelines and evaluation.

Zaghouani, W., & Charfi, A. (2018a). Arap-tweet: A large multi-dialect twitter corpus for gender, age, and language variety identification.

Zaghouani, W., Diab, M., Mansouri, A., Pradhan, S., & Palmer, M. (2010). The revised Arabic propbank.

Zaghouani, W., Habash, N., Obeid, O., Mohit, B., Bouamor, H., & Oflazer, K. (2016). Annotation Guidelines and Framework for Arabic Machine Translation Post-Edited Corpus.

Masking Explicit Pro-Con Expressions for Development of a Stance Classification Dataset on Assembly Minutes

Tomoyosi Akiba[1], **Gato Yuki**[1] , **Yasutomo Kimura**[2], **Yuzu Uchida**[3], **Keiichi Takamaru**[4]

[1]Toyohashi University of Technology
1-1 Hibarigaoka, Tempaku-cho, Toyohashi-shi, Aichi 441-8580 JAPAN
akiba@cs.tut.ac.jp, gato.yuki.am@tut.jp

[2] Otaru University of Commerce / RIKEN AIP
kimura@res.otaru-uc.ac.jp

[3] Hokkai-Gakuen University
yuzu@hgu.jp

[4] Utsunomiya Kyowa University
takamaru@kyowa-u.ac.jp

Abstract

In this paper, a new dataset for Stance Classification based on assembly minutes is introduced. We develop it by using publicity available minutes taken from diverse Japanese local governments including prefectural, city, and town assemblies. In order to make the task to predict a stance from content of a politician's utterance without explicit stance expressions, predefined words that directly convey the speaker's stance in the utterance are replaced by a special token. Those masked words are also used to assign a golden label, either agreement or disagreement, to the utterance. Finally, we constructed total 15,018 instances automatically from 47 Japanese local governments. The dataset is used in the shared Stance Classification task evaluated in the NTCIR-17 QA-Lab-PoliInfo-4, and is now publicity available. Since the construction method of the dataset is automatic, we can still apply it to obtain more instances from the other Japanese local governments.

Keywords: Stance Classification, Assembly Minutes, Automatic Data Construction

1. Introduction

In the recent development of electronic society, many public documents have been released in a digital format. Among them, assembly minutes are one of the most important documents for our society, since they contain many crucial congress decisions that impact on our daily life. Nevertheless, assembly minutes in themselves released by a various scale of governments (i.e., diet, congress, prefectural assembly, city and town councils) are undoubtedly difficult-to-read documents for human. Therefore, development of systems to exploit such documents for human with the help of NLP technology can be addressed as an urget issue.

An automatic analysis of text contents is a family of research problems including sentiment analysis (opinion mining), emotion recognition, argument mining (reason identification), sarcasm/irony detection, veracity and rumour detection, and fake new detection. Stance Classification is a recent member of them. The most common definition of Stance Classification is a task that identifies the standpoint of the producer of a piece of text towards a given target (Küçük and Can, 2020). We follow that definition in this paper.

The dataset on Stance Classification developed so far is mostly on online debate posts written in English. Indeed, the first competition on Stance Classification was carried out on microblog posts in English for a small number of pre-defined targets (Mohammad et al., 2016). Another common text type used for Stance Classification datasets is news texts (Ferreira and Vlachos, 2016). Recently, Barriere et al. (2022a), and their subsequent works (Barriere et al., 2022b; Barriere and Balahur, 2023), presented a new dataset of online debates. However, Stance Classification on assemblies had not been investigated for a long time since an earlier work in 2006 (Thomas et al., 2006). We believe that Stance Classification is indispensable for the analysis of assembly minutes since knowing the standpoint of each politician is one of the most basic functions to understand the debate conducted in them.

Based on the above, in 2020, an Stance Classification competition on Japanese assembly min-

Proceedings of the Second Workshop on Natural Language Processing for Political Sciences (PoliticalNLP 2024), pages 33–38

utes was carried out as a subtask of the NTCIR-15 QA-Lab-PoliInfo-2 (Kimura et al., 2020). The target text of the task was the assembly minutes of the Tokyo Metropolitan Assembly as a whole. A system participating in the task was given the minutes, a list of topics (agendas) discussed in it, a list of politicians participated in the discussion, and a political denomination list and was requested to classify each denomination's stance into two categories (agreement or disagreement) for each agenda. Through the evaluation, the task organizers and the participants of the task found that members of an assembly tend to state their stance on a given topic explicitly at the beginning of their speech. All participant's systems successfully exploited that surface text and achieved relatively good performance on the task.

Taking a lesson from the last Stance Classification task in the NTCIR-15 QA-Lab-PoliInfo-2, we designed a new Stance Classification task to identify the politician's stance from the content of their utterance without any explicit stance expression. Figure 1 illustrates our new Stance Classification task briefly. Since the original utterance includes an explicit stance expression '反対' (disagreement), it is rather straightforward to classify it into disagreement. Therefore, we replace such explicit expressions with a special token as shown in the masked utterance in the middle of the figure. Even without the expression '反対' (disagreement), we can deduce its stance of opposition from the underlined part.

In this paper, we report our effort of developing the new dataset of Stance Classification on Japanese local assembly minutes. We developed an automatic method of data construction and finally collected 4,324 and 10,694 instances for the dry run and the formal run evaluations of the PoliInfo-2 Stance Classification task, respectively, from total 47 Japanese local governments of various city and town assemblies.

The rest of the paper is organized as follows. In Section 2, we summarize the related work on Stance Classification and the previous shared task evaluated in the NTCIR-15 QA-Lab-PoliInfo-2. In Section 3, the task design of our Stance Classification task is explained. In Section 4, the detailed methods of developing our dataset are described. In Section 5, the evaluation results on the shared task, NTCIR-17 QA-Lab-PoliInfo-4, are presented. Finally, in Section 6, the outcome of our work is summarized.

2. Related Work

2.1. Stance Detection

Stance Classification (also known as Stance Detection) is a task that identifies the standpoint of the producer of a piece of text towards a given target (Küçük and Can, 2020). The earliest competition on the task is SemEval-2016 shared task on Twitter stance detection (Mohammad et al., 2016). After that, various datasets have been created mainly on social media(Xu et al., 2016; Taulé et al., 2017; Glandt et al., 2021). Until the previous NTCIR-15 QA-Lab-PoliInfo-2 Stance Classification task evaluated in 2020 (Kimura et al., 2020), stance classification on assemblies had not been investigated for a long time since an earlier work in 2006 (Thomas et al., 2006).

2.2. NTCIR-15 QA-Lab-PoliInfo-2

The NTCIR-15 QA-Lab-PoliInfo-2 Stance Classification task aims at estimating politician ' s position from politician ' s utterances (Kimura et al., 2020). A system participating in the task estimates the stances of political parties from the utterances of the members of the Tokyo Metropolitan Assembly. Given the Tokyo Metropolitan Assembly, topics (agenda), member ' s list and political denomination list, the systems classify their stance into two categories (agreement or disagreement) for each agenda. Five teams were participated in the formal run evaluation. All of them exploited the explicit stance expression appeared at the beginning of the politician's speeches to achieve their good performances.

3. Task Design

Since we constructed a dataset used for the Stance Classification-2 task evaluated at the shared task, the NTCIR-15 QA-Lab-Poliinfo-4, we will firstly describe our task design of the Stance Classification-2 task.

The Stance Classification task aims at estimating politician's position from her/his utterances. Taking a lesson from the last Stance Classification task evaluated at the NTCIR-15 QA-Lab-Poliinfo-2, we revisit it by taking into account the following two aspects. Firstly, we redesign the classification task itself. In the last task, the information source of the classification was assembly minutes as a whole. The task organizers found that members of an assembly tend to state their stance on a given topic explicitly at the beginning of their speech. While most of the participants successfully exploited that to achieve good performance, the use of such superficial expression does not well matched with our

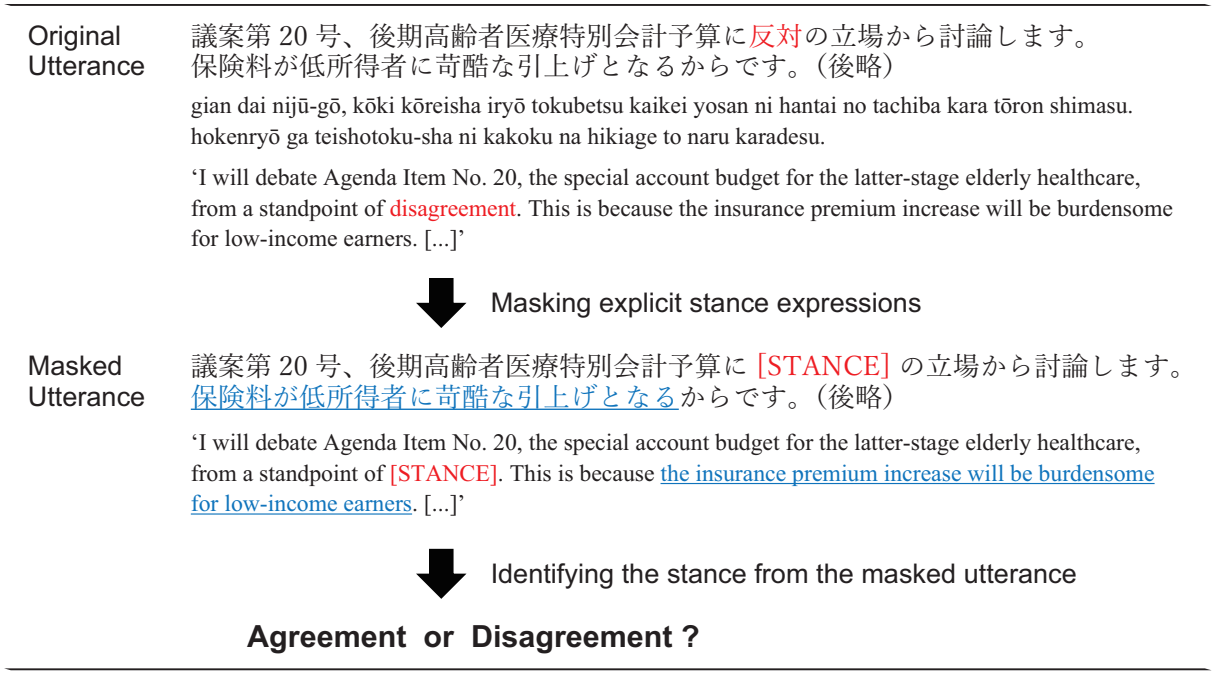


Figure 1: Workflow of the stance classification-2 task

purpose, i.e., estimating politician's position from the contents of her/his utterance. Therefore, in the new Stance Classification-2 task, we focus on a classification of members' opinion about a given topic without any explicit statement on their stance. Secondly, we extend the target minutes to several local governments in Japan other than Tokyo Metropolitan Assembly.

In order to ensure that given utterance does not include explicit expressions about neither agreement nor disagreement, we adopt a simple manipulation on it. We found that only few Japanese words play a critical role of expressing stance of speakers and that those words can be used exchangeabely without losing grammatical correctness. Therefore, we found a simple replacement of such words with a pre-defined mask token serves our purpose. We chose two Japanese words, '賛成' (agreement) and '反対' (disagreement), for such words. Note that those words can be used as verbs by postfixing a Japanese function word 'する' as seen in '賛成する' (agree) and '反対する' (disagree), so the replacement method also works for such cases.

The category label often used in conventional Stance Classification task is either Favor, Against, or Neither (Mohammad et al., 2016). On the other hand, that used in our Stance Classification-2 task is either Agree or Disagree. That is because, in a political assembly, politicians must have a unambiguous decision on a given topic since they finally take a vote on it.

In summary, the Stance Classification-2 task is defined as follows: Given a masked utterance of a politician associated with a topic (agenda), participant's systems are requested to classify it into two categories (agreement or disagreement).

Table 1: Selected Local Assemblies

prefecture	#city	#town
Aichi	9	5
Hokkaido	4	1
Saitama	18	9
Fukuoka	1	0
TOTAL	32	15

4. Data Construction

4.1. Selection of local governments

Japan has diverse local governments, each of which has its own assembly, e.g. prefectural assembly, city council, and town council. We found that their styles of discussion are divided into two groups. One of them is that a topic is discussed separately, on each of which representative politicians express their own opinions. The other is that multiple topics are discussed at the same time so a representative politician express her/his opinions on them continuously. We focused on the former group since it was easier to extract a politician's utterance associated with a specific topic than the latter. Finally, we selected 47 local assemblies from Aichi, Hokkaido, Saitama, and Fukuoka prefectures. Table 1 shows the details.

4.2. Extraction and Labeling of Politician's Utterances

We extracted politician's utterances on the last day of a series of a regular meeting, in which they take a vote on a given topic so they should have decided their position clearly. Since each utterance of an assembly minutes is associated with a speaker label and a chairperson presides at the order of discussed topics and speakers, consecutive utterances spoken by a specific politician and directed to a specific topic are unambiguously extracted.

In order to ensure that the utterances have no explicit expression about speaker's stance, some selected tokens are replaced with a special token `[STANCE]`. We chose '賛成' (agreement) and '反対' (disagreement) for such selected tokens. At the same time, we utilize those tokens to assign a golden label to the utterance by using the following heuristic rules.

1. If the selected tokens that appear in an utterance are all the same, namely either all agreement or all disagreement, then assign the golden label accordingly.

2. Otherwise, if the utterance includes some formulaic expression that clearly express the speaker's stance, then assign the golden label accordingly. The regular expression pattern used for the formulaic expressions are:

   (賛成|反対)((の|をする) 立場 (で|から))?(討論を)?(させていただき|いたし|申し上げ|し)ます。
   (I will argue from a position of (support|opposition).)

3. Otherwise, discard the utterance from the dataset.

Through our preliminary experiments, we found that method seldomly assigned incorrect labels.

4.3. Dataset Details

We distributed two separate CSV files for training and test data, whose data fields are shown in Table 2. In the test data, 'stance' field is left blank and participant's systems are requested to fill it with either 'agreement' or 'disagreement'. For the dry run, we released 3,898 and 426 instances for training and test data, respectively, which were constructed from 19 local governments in Aichi and Hokkaido prefectures. For the training data of the formal run, we released 8,534 instances constructed from 26 local governments of Saitama prefecture. For the test data of the formal run, we released 2,160 instances from 27 (same 26 and one more) local governments of Saitama prefecture and 80 instances from (hidden) one local government of Fukuoka prefecture.

Table 2: Data fields of the stance classification 2 task

Field name	Explanation
id	Question ID (Japanese local government ID and serial number)
prefecture	Name of the prefecture
assembly	Name of the local government
meeting	Name and serial number of the regular meeting
date	Date of the meeting
speaker	Speaker name of the utterance
utterance	An utterance by an politician whose explicit tokens are replaced with `[STANCE]`
target	topic of the utterance
stance	'Agreement' or 'Disagreement'

In addition to the regular training data above, we also released their unmasked version, in which the texts in the 'utterance' field are not masked, i.e., the selected explicit tokens are not replaced with `[STANCE]` but are left unchanged, hoping participants may use it for their system development.

Table 3: Statistics of data

	Training Data	Test Data
Dry Run	3,898	426
Formal Run	8,534	2,160

5. Evaluation

5.1. NTCIR-17 QA-Lab-PoliInfo-4

Table 4: Summary of participants' methods and results

Team	Pre-Trained Model	Accuracy
KIS	LUKE	0.9728
ISLab	GPT-3	0.9326
AKBL	RoBERTa	0.9308

The dry run evaluation took place from March 6th to July 3rd in 2023. The formal run evaluation took place from July 4th to 15th in 2023. During those days, task participants were allowed to submit their classification results to our leader board system. In the end of the formal run, we had three active task participant teams. All of them employed pre-trained language models for the basis of their classifiers. Their pre-trained language models and evaluation results are summa-

rized in Table 4. The detail of the task is found in (Ogawa et al., 2023).

6. Conclusion

This paper described about a new dataset for Stance Classification based on assembly minutes. It was developed by using publicity available minutes taken from diverse Japanese local governments including city and town assemblies from several prefectures. For each politician's utterance in the dataset, the words of expressing either agreement or disagreement were masked by a special token, in order to make the task to predict a stance from content of a politician's utterance. Those masked words were also used to assign a golden label.

Finally, we constructed total 15,018 instances automatically from 47 Japanese local governments selected from four prefectures. Using the dataset, the shared task of Stance Classification was evaluated in the NTCIR-17 QA-Lab-PoliInfo-4. The dataset is now publicity available[1].

Since the construction method of the dataset is automatic, we will apply it to obtain more instances from the other Japanese local governments in our future work.

7. References

Valentin Barriere and Alexandra Balahur. 2023. Multilingual multi-target stance recognition in online public consultations. *Mathematics*, 11(9).

Valentin Barriere, Alexandra Balahur, and Brian Ravenet. 2022a. Debating Europe: A multilingual multi-target stance classification dataset of online debates. In *Proceedings of the LREC 2022 workshop on Natural Language Processing for Political Sciences*, Marseille, France. European Language Resources Association.

Valentin Barriere, Guillaume Guillaume Jacquet, and Leo Hemamou. 2022b. CoFE: A new dataset of intra-multilingual multi-target stance classification from an online European participatory democracy platform. 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 2: Short Papers)*, Online only. Association for Computational Linguistics.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California. Association for Computational Linguistics.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in COVID-19 tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online. Association for Computational Linguistics.

Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Keiichi Takamaru, Madoka Ishioroshi, Teruko Mitamura, Masaharu Yoshioka, Tomoyoshi Akiba, Yasuhiro Ogawa, Minoru Sasaki, Kenichi Yokote, Tatsunori Mori, Kenji Araki, Satoshi Sekine, and Noriko Kando. 2020. Overview of the NTCIR-15 QA Lab-PoliInfo task. *Proceedings of The 15th NTCIR Conference*.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. 53(1).

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41.

Yasuhiro Ogawa, Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Keiichi Takamaru, Kazuma Kadowaki, Tomoyoshi Akiba, Minoru Sasaki, Akio Kobayashi, Masaharu Yoshioka, Tatsunori Mori, Kenji Araki, and Teruko Mitamura. 2023. Overview of the NTCIR-17 QA Lab-PoliInfo task. *Proceedings of The 17th NTCIR Conference*.

Mariona Taulé, M Antonia Martí, Francisco M Rangel, Paolo Rosso, Cristina Bosco, Viviana Patti, et al. 2017. Overview of the task on stance and gender detection in tweets on catalan independence at ibereval 2017. In *CEUR Workshop Proceedings*, volume 1881, pages 157–177. CEUR-WS.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. *arXiv preprint cs/0607062*.

Ruifeng Xu, Yu Zhou, Dongyin Wu, Lin Gui, Jiachen Du, and Yun Xue. 2016. Overview of

[1] https://github.com/poliinfo4/PoliInfo4-FormalRun-Stance-Classification-2

nlpcc shared task 4: Stance detection in chinese microblogs. In *Natural Language Understanding and Intelligent Applications: 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016, Kunming, China, December 2–6, 2016, Proceedings 24*, pages 907–916. Springer.

Analysing Pathos in User-Generated Argumentative Text

Natalia Evgrafova, Véronique Hoste, Els Lefever
LT3, Ghent University, Belgium
Groot-Brittanniëlaan 45, 9000 Ghent
{natalia.evgrafova, veronique.hoste, els.lefever}@ugent.be

Abstract

While persuasion has been extensively examined in the context of politicians' speeches, there exists a notable gap in the understanding of the pathos role in user-generated argumentation. This paper presents an exploratory study into the pathos dimension of user-generated arguments and formulates ideas on how pathos could be incorporated in argument mining. Using existing sentiment and emotion detection tools, this research aims to obtain insights into the role of emotion in argumentative public discussion on controversial topics, explores the connection between sentiment and stance, and detects frequent emotion-related words for a given topic.

Keywords: argument mining, sentiment analysis, emotion detection, social media, political discourse

1. Pathos in Political Argument Mining

An essential aspect of political activity is persuasive communication (Windisch, 2008). According to Wolton (1989), political communication is a platform where politicians, journalists, and the public (through opinion polls) openly express their views on politics. Windisch (2008) rightly argues that public opinion today is expressed through a variety of communication channels available. This trend has led to the development of both automated and non-automated methods for public opinion collection and analysis. With the advances in big data, tools such as sentiment analysis, opinion and argument mining have been developed to understand and predict public attitudes towards various entities, from products or films to governmental initiatives and salient social problems.

Persuasive communication in politics implies effective argumentation, achieved within logos, pathos, and ethos dimensions (Cardoso et al., 2023). The logos dimension is associated with the logical structure of arguments, the pathos dimension is related to appeals to emotion, and ethos is concerned with credibility and appeals to authority (Cardoso et al., 2023; Habernal and Gurevych, 2017).

In natural language processing, it is the field of argument mining that aims to automatically extract, analyse and understand arguments from natural language text. Within argument mining, researchers have been developing methods to classify argumentative and non-argumentative spans of text, detect topics, aspects, stances, and other argument components, and generate high-quality arguments (Cabrio and Villata, 2018). While their work has predominantly focused on the logos dimension, with a particular success for argumentative essays and other well-structured text, analysing the pathos dimension has not been as thorough despite the fact that it plays an important role in the social media argumentative discourse, especially political discussions. Such discussions often include emotional and metaphoric language that cannot be properly analysed within the logos dimension (Habernal and Gurevych, 2017). Moreover, attempts to analyse these arguments are associated with challenges related to overlaps of sentiment analysis and argument mining (Cabrio and Villata, 2018), and low inter-annotator agreement for emotional components of arguments (Habernal and Gurevych, 2017).

To the best of our knowledge, the exploration of the pathos dimension of argumentation, including the appeal to emotion, have been primarily dealt with within fallacy detection (Goffredo et al., 2023; Vijayaraghavan and Vosoughi, 2022; Sahai et al., 2021). In argumentation, logical reasoning is considered to be more legitimate than emotional language (Duckett, 2020). However, an appeal to emotion does not necessarily mean fallacious argumentation (Walton, 2005; Duckett, 2020), and its usage could be justified when it comes to value-contested debates such as assisted dying, abortion, war, independence (Duckett, 2020).

In Walton (1992), we read that certain types of emotional appeals "are very powerful as arguments in themselves", though there is always a chance they could be fallacious, namely, irrelevant or logically weak, but that is not always the case and, more importantly, does not always limit their effect. It is often through emotional language that users express their beliefs, values, and moral motivations. This is why we argue that confining argument mining solely to the logos dimension and reserving the

Proceedings of the Second Workshop on Natural Language Processing for Political Sciences (PoliticalNLP 2024), pages 39–44

pathos for fallacy detection may prove overly restrictive. Simply marking argumentation as fallacious might not substantially improve our understanding of the prevalent reasons for taking one stance or the other.

An effort to include the pathos dimension in annotations was made by Habernal and Gurevych (2017) in their study on argument mining in user-generated web discourse. In this study, the corpus included documents that were retrieved from heterogeneous web resources (comments on articles, forum and blog posts). 6% of documents were purely emotional without logical backing and could not be analysed in terms of their logical structure. Though in some cases claims and premises could be identified, persuasiveness was achieved through emotion. Following the given annotation guidelines, annotators had to classify arguments as "appeal to emotion" in case the argumentation relied on figurative language or obvious exaggerations. The task posed a significant difficulty reflected in Krippendorf's agreement of only $\alpha U = 0.30$. Consequently, the authors chose to focus on the logos dimension. The Internet Argument Corpus (IAC) (Walker et al., 2012) included annotation for Fact/Emotion-based arguments with relatively low agreement results $\alpha U = 0.32$. The low inter-annotator agreement proves that new approaches should be developed for incorporating the pathos dimension into argument mining. Logical and emotional components in arguments are intertwined (van Eemeren and van Haaften, 2023), expressed in different degrees and supplement each other for the purpose of persuasiveness, making emotion an integral part of public reasoning (Stucki and Sager, 2018) that should not be disregarded.

In this work we focused on an exploratory pathos analysis of argumentative text for the task of argument mining. The contributions of the paper are the following: (1) analysing the relationship between sentiment and stance, (2) comparing the results of a manual analysis of the emotional components of arguments on a given topic with an automated extraction of emotion-related words, and (3) suggesting ways to incorporate the pathos dimension for the argument mining task.

2. The Datasets

For the analysis of sentiment, emotional words, and the connection between sentiment and stance, we selected two datasets consisting of user-generated arguments on contentious topics with stance annotations. We prioritised stance annotations over sentiment annotations as stance is more difficult to detect automatically, but we could automatically annotate the argument sentiment with sufficient accuracy. We chose the Webis *args.me* corpus (Ajjour et al., 2019) containing 387,606 arguments from debate portals and the IBM ArgKP-2021 dataset (Friedman et al., 2021) of crowdsourced arguments — based on the ArgKP dataset (Bar-Haim et al., 2020) and Gretz et al. (2019). As some of the topics in the Webis corpus contained very few comments, we decided on the 30 most commented topics from the corpus and deleted very short comments (up to 10 words) such as "I win" that were part of users' communication on the platform, which resulted in 8902 comments. From the IBM ArgKP-2021 dataset we kept 7238 unique arguments on 31 topics.

3. Corpora Analysis

3.1. Sentiment and Stance

The relationship between sentiment and stance is complex. One of the hypotheses could be that the sentiment is more positive in PRO stances and more negative in CON stances. This could be true for certain datasets and explain why BERT-based models tend to rely on sentiment words for stance prediction (Trautmann, 2020). However, the same arguments might be attacking a certain topic or aspect and support the other, regardless of their sentiment; for example, an argument attacking coal energy might be supporting wind energy (Daxenberger et al., 2020). Understanding how sentiment is connected with stance might not only provide insights into the pathos dimension, but also help design corpora in such a way as to minimise errors in machine-learning.

The first task that we addressed was the analysis of the sentiment and stance distribution in the chosen corpora. To compensate for the lack of sentiment annotations, the arguments in both corpora were automatically annotated for positive, neutral, or negative overall sentiment using two existing transformer-based models for sentiment analysis with reasonable recall scores. When selecting a model, our aim was to ensure that it was trained and fine-tuned on internet user-generated texts. We prioritised models based on tweets due to their closer resemblance to our dataset: tweets are user-generated, vary in length, and incorporate colloquial and emotional language. The first model we used was the recent version of the fine-tuned twitter-roberta-base-sentiment-latest (Camacho-Collados et al., 2022). The second model we applied was the fined-tuned pysentimiento bertweet-base-sentiment-analysis model (Pérez et al., 2023). The comparison of the resulting sentiment labels from the two models showed an overlap of about 78% in both corpora, which meant that the majority of arguments were correctly labelled in terms of their sentiment. The further exploration was based on

the labels from the twitter-roberta-base-sentiment-latest as it allowed for longer texts (512 tokens compared to 128 in pysentimiento).

The preliminary analysis of the Webis dataset revealed a larger proportion of neutral and positive texts in PRO arguments (61,3%) compared to CON (50,0%).

In the IBM corpus, the proportion of negative arguments for PRO was higher (55.1% compared to 47.2% in CON), as was the proportion of positive arguments (14.4% compared to 11.5% in CON). Conversely, there were more arguments with a neutral sentiment in CON (41.3% compared to 30.5% in PRO), see Fig. 1. A qualitative analysis showed that for this corpus, the larger proportion of negative sentiment in PRO could be explained by the fact that many topics, which are major claims in this corpus, are formulated with a negative framing, for instance, “Assisted suicide should be criminal offence”, “We should ban human cloning”, “Home schooling should be banned”. These claims imply a negative stance towards the main topic (assisted suicide, home schooling, human cloning) making arguments that are against main topics actually belonging to the PRO category if the topic is positively framed.

To check if swapping the stance labelling for such topics results in major changes in the sentiment distribution, we studied 31 topics of the IBM dataset and manually changed the stances for 18 topics that were framed negatively.

The results for the modified IBM corpus (see Fig. 2) revealed a much bigger proportion of positive and neutral sentiment in PRO and a substantial increase in the negative sentiment in CON. This indicated that topic framing influenced the distribution of sentiment across stances.

In the 30 topics of the Webis dataset there were only two negatively framed topics “Abortion should be illegal” and "Gun Control" with 323 comments, which could not result in much change. Among other topics, there were "Gun rights", "Abortion", "Gay marriage", "Euthanasia" that implied a positive claim even though some of them were expressed in one word only, for example, "Abortion" could be extended to "Abortion should be allowed" without causing changes in the stances of the arguments. However, it should be noted that arguments in this corpus included quotations of the opposing position, and there were some non-argumentative user-interaction comments, which could also influence the sentiment distribution.

To conclude this section, from certain datasets a model can learn to rely too much on sentiment for stance prediction, but datasets can be modified to reduce errors. One of the ways to decrease the impact of sentiment on stance is to check the sentiment and stance distribution in the training corpus and ensure the balance. Another way would be to conduct training that involves positively and negatively framed similar topics (ex: "Abortion should be allowed", "Abortion should be banned", or "Abortion rights", "Abortion ban") for the same arguments, which could yield more robust results. On the whole, sentiment is intricately connected with stance and is highly influenced by a topic and its framing, certain checks and dataset modifications could be used to lower the chances of short-cut machine learning.

3.2. Emotion Words in Arguments

To get the first insights into the emotional dimension of arguments, we chose the "Death penalty" ("Capital punishment") topic to analyse in both corpora. First, we conducted manual analysis on a sample of arguments to identify emotional components. Next, we automatically annotated arguments with the NRClex 4.0 affect generator[1] based on the NLTK library’s WordNet synonym sets (Bird et al., 2009) and the NRC lexicon (Mohammad and Turney, 2013). The final labels for emotion included emotion-related words from each argument and a list of emotions associated with them. This enabled comparison with the results of the manual analysis.

The qualitative analysis of 171 arguments (90 PRO, 81 CON) for the topic of "death penalty" in the Webis corpus showed that in PRO arguments the prevalent emotional language was used to describe criminals that were “dangerous”, “heartless”, “cold-blooded”, often mentioning paedophiles, that “deserved” this punishment for the “heinous”, “violent” criminal act and “awful", "horrible" things they committed. This was contrasted with “innocent” victims “condemned to a terrible life” who “deserve” “true”, “proportional” justice. The Old Testament quotations and especially the "eye for eye" principle were referred to in the context of the punishment being "justified" and serving as an "efficient deterrent" that "inspires" "fear" into criminals. The concept of capital punishment was described as a means of "protection" of the "innocent" that brings "peace", "solace" and "closure", and "saves" other people.

The CON arguments described capital punishment as "a murder", "cruel", "outdated", "barbaric", "racist", "sexist, "unnecessary", "expensive", "hypocritical", they included appeals to "forgiveness", a chance for criminals to "repent" and used "innocent" to refer to the wrongly executed people.

In the CON arguments, there were also frequent mentions of "better", "more efficient", "other" ways to punish criminals, as well as references to "equality", "human rights",and questioning if death penalty is a "good" deterrent, describing life in prison as a "greater punishment".

[1](C) 2019 Mark M. Bailey, PhD

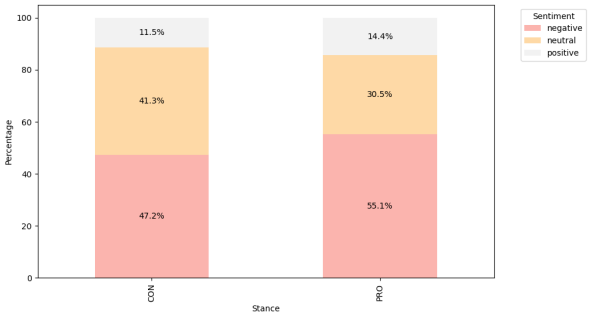


Figure 1: IBM sentiment distribution by stance

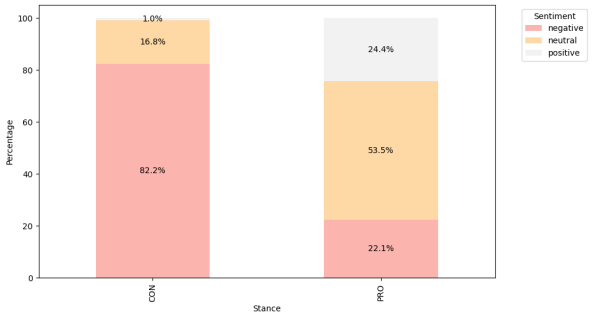


Figure 2: inverted IBM sentiment distribution by stance

As observed by Walton (2005), some argumentative discussions include an argument about how to define key concepts. This was also seen in our data, e.g. "Death penalty is a murder", as opposed to "Death penalty is a deterrent". Such *definitions* are highly important in ethically controversial debates, and they tend to differ in terms of the emotional spin for opposing parties (Walton, 2005).

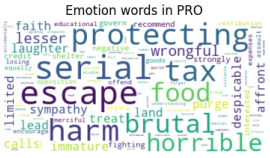
The next step was to compare these results with what a simple emotional words detection could yield depending on arguments' stance. For this purpose, we relied on the NRClex 4.0 affect generator to extract only emotion-related words from all the comments. Subsequently, we segregated these words based on the stance of the arguments within the selected topic, creating unique sets for both 'PRO' and 'CON' stances. During this process, words exclusive to each stance were identified, with any overlapping words removed. After that, we counted the frequencies of these unique words and used these counts to generate word clouds that represented the most common emotion-related words for each stance.

The word clouds for the 171 arguments from the Webis corpus that had been previously manually analysed showed the prevalence of "protecting" people, the risk of prisoners' "escape" and references to "brutal", "horrible" things for PRO and a higher frequency of religious references to "hell", "repent", and "spirit" in the CON category (see Fig. 3).

For comparison, the world clouds for the "capital punishment" topic from the IBM corpus featured "violence" and "ineffective" as the most frequent emotional words for CON and the high frequency of "heinous" and "deserve" for PRO (see Fig. 4).

Overall, this exploration provided general insights into emotional language associated with stances. Nevertheless, some results were not easy to interpret without knowing what they referred to or a deeper knowledge of the context.

Based on these observations, we suggest considering the following in order to detect pathos in argumentative text: (1) the emotional components of arguments are defining particular aspects ("heinous crime", "innocent victim"). A more fine-grained comparison of emotional words by aspect could bring about more insightful results. Aspects can also be emotionally loaded and expressed by a variety of lexical means (e.g. "Solace"/"Peace"/"Closure" for crime victims); (2) apart from retrieving emotional words by aspect, pathos could be further explored through extraction of key concepts persuasive definitions (e.g. "Death penalty is a murder") as they often contain emotional words that differ for PRO and CON stances (Walton, 2005); (3) for comparison of PRO and CON, the topic should be clearly defined and be controversial, its framing, negative or positive, should be taken into account.



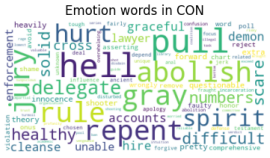
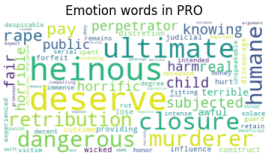


Figure 3: "Death penalty", 171 comments, Webis



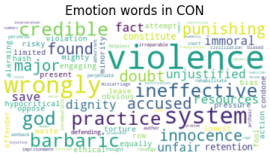


Figure 4: "Capital punishment", 236 comments, IBM

4. Conclusion and Future Work

This paper explored the possibility of incorporating the pathos dimension of argumentation for the task of argument mining. The IBM ArgKP-2021 and the Webis args.me corpora were used for the analysis of the relation between sentiment and stance, and emotional words detection. A part of the Webis corpus was manually analysed to compare the results and develop ideas for automation of the pathos dimension analysis.

The automatic detection of emotional words

based on a lexicon-based approach provided the first insights into the pathos dimension of arguments based on their stance, however, certain results were difficult to interpret without contextual information and understanding which aspects those words referred to.

A first avenue for future work is the extraction of definitions and emotional words associated with specific aspects in the argumentative text across diverse topics. Another important direction consists in developing effective annotation guidelines and creating pathos-annotated corpora of the argumentative texts. Finally, we aim to explore various pipelines to automate stance-dependent pathos analysis in argumentative texts on contentious topics.

Gaining deeper insights into the pathos dimension of arguments in political social media text can shed light on the role of emotion across controversial topics and in forming public opinions. Further developments could deepen our comprehension of what convinces the public, which stories and interpretations get spread in different languages, and how the public responds to these stories, what is reproduced and challenged.

Acknowledgements

This work was supported by the Research Foundation – Flanders (FWO) under grant FWO.OPR.2023.0004.01 (G019823N).

5. Bibliographical References

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020. From arguments to key points: Towards automatic argument summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4029–4039. Association for Computational Linguistics.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: a data-driven analysis. In *27th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5427–5433. International Joint Conferences on Artificial Intelligence Organization.

Jose Camacho-Collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa Anke, Fangyu Liu, and Eugenio Martínez Cámara. 2022. TweetNLP: Cutting-edge natural language processing for social media. pages 38–49. Association for Computational Linguistics.

Henrique Lopes Cardoso, Rui Sousa-Silva, Paula Carvalho, and Bruno Martins. 2023. Argumentation models and their use in corpus annotation: Practice, prospects, and challenges. *Natural Language Engineering*, 29(4):1150–1187.

Johannes Daxenberger, Benjamin Schiller, Chris Stahlhut, Erik Kaiser, and Iryna Gurevych. 2020. ArgumenText: Argument classification and clustering in a generalized search scenario. *Datenbank-Spektrum*, 20(2):115–121.

Stephen Duckett. 2020. Pathos, death talk and palliative care in the assisted dying debate in Victoria, Australia. *Mortality*, 25(2):151–166.

Pierpaolo Goffredo, Mariana Chaves, Serena Villata, and Elena Cabrio. 2023. Argument-based detection and classification of fallacies in political debates. pages 11101–11112. Association for Computational Linguistics.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2019. A large-scale dataset for argument quality ranking: Construction and analysis.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Juan Manuel Pérez, Mariela Rajngewerc, Juan Carlos Giudici, Damián A. Furman, Franco Luque, Laura Alonso Alemany, and María Vanina Martínez. 2023. pysentimiento: A Python toolkit for opinion mining and social NLP tasks.

Saumya Sahai, Oana Balalau, and Roxana Horincar. 2021. Breaking down the invisible wall of informal fallacies in online discussions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 644–657. Association for Computational Linguistics.

Iris Stucki and Fritz Sager. 2018. Aristotelian framing: logos, ethos, pathos and the use of evidence in policy frames. *Policy Sciences*, 51(3):373–385.

Dietrich Trautmann. 2020. Aspect-based argument mining. In *Proceedings of the 7th Workshop on Argument Mining*, pages 41–52. Association for Computational Linguistics.

Frans H. van Eemeren and Ton van Haaften. 2023. The making of argumentation theory: A pragma-dialectical view. *Argumentation 2023*, 37(3):341–376.

Prashanth Vijayaraghavan and Soroush Vosoughi. 2022. TWEETSPIN: Fine-grained propaganda detection in social media using multi-view representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3433–3448. Association for Computational Linguistics.

Marilyn Walker, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 812–817. European Language Resources Association.

Douglas Walton. 1992. *The Place of Emotion in Argument*. Penn State University Press.

Douglas Walton. 2005. *Fundamentals of Critical Argumentation*. Cambridge University Press.

Uli Windisch. 2008. Daily political communication and argumentation in direct democracy: advocates and opponents of nuclear energy. *Discourse Society*, 19(1):85–98.

Dominique Wolton. 1989. La communication politique: construction d'un modele. *Hermès, La Revue*, 4(1):27–42.

6. Language Resource References

Yamen Ajjour and Henning Wachsmuth and Johannes Kiesel and Martin Potthast and Matthias Hagen and Benno Stein. 2019. *Data Acquisition for Argument Search: The args.me corpus*. Springer. PID https://zenodo.org/records/3274636.

Bird, Steven and Klein, Ewan and Loper, Edward. 2009. O'Reilly. PID https://www.nltk.org/nltk_data/.

Friedman, Roni and Dankin, Lena and Hou, Yufang and Aharonov, Ranit and Katz, Yoav and Slonim, Noam. 2021. *Overview of the 2021 Key Point Analysis Shared Task*. Association for Computational Linguistics. PID https://paperswithcode.com/dataset/argkp-2021.

Mohammad, Saif M. and Turney, Peter D. 2013. *Crowdsourcing a Word–Emotion Association Lexicon*. PID http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm.

Knowledge Graph Representation for Political Information Sources

Tinatin Osmonova, Alexey Tikhonov, Ivan P. Yamshchikov
OSCE Academy, Bishkek
Inworld.AI, Berlin
CAIRO, Technical University of Applied Sciences Würzburg-Schweinfurt (THWS), Würzburg
ivan.yamshchikov@thws.de

Abstract

With the rise of computational social science, many scholars utilize data analysis and natural language processing tools to analyze social media, news articles, and other accessible data sources for examining political and social discourse. Particularly, the study of the emergence of echo-chambers due to the dissemination of specific information has become a topic of interest in mixed methods research areas. In this paper, we analyze data collected from two news portals, Breitbart News (BN) and New York Times (NYT) to prove the hypothesis that the formation of echo-chambers can be partially explained on the level of an individual information consumption rather than a collective topology of individuals' social networks. Our research findings are presented through knowledge graphs, utilizing a dataset spanning 11.5 years gathered from BN and NYT media portals. We demonstrate that the application of knowledge representation techniques to the aforementioned news streams highlights, contrary to common assumptions, shows relative "internal" neutrality of both sources and polarizing attitude towards a small fraction of entities. Additionally, we argue that such characteristics in information sources lead to fundamental disparities in audience worldviews, potentially acting as a catalyst for the formation of echo-chambers.

Keywords: echo chambers, computational social science, knowledge representation

1. Introduction

A knowledge graph, also known as a semantic network, was initially introduced by C. Hoede and F.N. Stokman as a tool for representing the content of medical and sociological texts (Nurdiati and Hoede, 2008). Constructing increasingly larger graphs with the intent of accumulating knowledge was initially deemed to provide a resultant structure capable of operating as an expert system proficient in investigating causes and computing the consequences of certain decisions.

The concept of knowledge graph co-evolved with the rise of computational social science (Conte et al., 2012) and digital data analysis methods (Rogers, 2013). Access to open sources on the Internet has facilitated the measurement of the dynamics of political debates (Neuman et al., 2014). Platforms like Twitter and other microblogging services are widely utilized for studying and modeling social and political discourse (Graham et al., 2016), (Jungherr, 2014) , (Wang et al., 2018). Contemporary researchers even develop a conceptual framework for predicting the morality underlying political tweets(Johnson and Goldwasser, 2018). Moreover, knowledge graphs of fact-checked claims, such as ClaimsKG, have been designed. Such tools facilitate structured queries about truth values, authors, dates, journalistic reviews, and various types of metadata (Tchechmedjiev et al., 2019).

A significant group of studies, advocate usage of graphs for social, political, and business industry data, stating that "graphs greatly increases the clarity of presentation and makes it easier for a reader to understand the data being used"(Kastellec and Leoni, 2007) . Additionally, (Abu-Salih and Beheshti, 2021) explains that knowledge graphs serve as indispensable frameworks that underpin intelligent systems. This is achieved by extracting subtle semantic nuances from textual data sourced from a range of vocabularies and semantic repositories. In the past decade, there has been a notable increase in the examination of political discourse within social content in such a way. The authors discuss in detail the connection between political discussions and the language used in them (Chilton, 2004), (Parker, 2014). Furthermore, the literature examines opinion polarization (Banisch and Olbrich, 2019), attempts to characterize an intuition of the dynamics of the political debate (Yamshchikov and Rezagholi, 2019), and provides techniques for estimating them (Merz et al., 2016), (Subramanian et al., 2017), (Glavaš et al., 2017), (Subramanian et al., 2018) or (Rasov et al., 2020). The extensively employed data sources in studies centered on automated text classification for political discourse analysis involve Manifesto Database (Lehmann et al., 2017) and the proceedings of the European Parliament (Koehn, 2005).

The challenges arising in contemporary studies on observational and discourse analysis are the quality of data (Tweedie et al., 1994) and the credibility of data sources. It is crucial to apply statistical measures and tests to quantify the impact of poor data quality and bias on the results (Abu-Salih and Beheshti, 2021). However, quantifying such effects proves comprehensive in the realm of social sciences due to the numerous indigent properties of

Proceedings of the Second Workshop on Natural Language Processing for Political Sciences (PoliticalNLP 2024), pages 45–54

social datasets (Shah et al., 2015). One significant challenge is associated with the formation of so-called echo-chambers in social structures, which naturally obstruct the propagation of information, reinforcing disparities across various social strata (Goldie et al., 2014), (Colleoni et al., 2014), (Guo et al., 2015) or (Harris and Harrigan, 2015). Addressing the credibility of sources, the phenomenon of fake news draws constant attention from media outlets and researchers. According to (Anderson and Auxier, 2020), 55% of online social network users believe they are accurately informed about recent political updates by the media. Consequently, misleading information and false news have the potential to shape certain beliefs and human behaviors. As a solution, several studies (Allcott and Gentzkow, 2017), (Shu et al., 2017) or (Lazer et al., 2018) analyze and propose methods to enhance the quality of information. Additionally, these studies imply the existence of a certain ground truth that could be universally accepted.

Taking existing knowledge and challenges into account, in this work, we study the issue of news representation from a data analysis perspective. We construct two datasets comprising news articles from "alt-right" and "liberal" news platforms, denoted as Breitbart News (BN) and the New York Times (NYT), spanning 11.5 consecutive years (from 2008 to Fall 2019). We demonstrate that information disparities between these news sources are fundamental regardless of the social structures that encapsulate the readers of the aforementioned outlets. Upon analyzing the findings, we assert that one has to take into consideration these disparities, since they signify fundamental differences in the foundational data that shapes the perspectives, beliefs, and, ultimately, the behavior of readers. Simply put, even if we had no social media information disparities by various news sources could contribute to echo-chamber formation.

2. Data and Methodology

We have parsed two news sites Breitbart News[1] that could be generally associated with the "alt-right" political views and the New York Times[2] associated with "liberal" political views. The choice of these two media platforms was arbitrary to a certain extent. We parsed all news presented on both platforms in the period from 2008 till the fall of 2019. Using the texts of the news as input data we built an information extraction pipeline aimed to reconstruct a form of knowledge graph out of the news texts. To do that we have used state of the art open information extraction (Stanovsky et al., 2018) and named entity recognition (Peters et al., 2017) tools of AllenNLP[3]. The outputs of both models are noisy, so in order to stabilize the resulting signal we came up with the heuristics for substring-matching. We used only ARG0 and ARG1 items of open information extractor and all entities of named entity recognition to extract the most useful objects of the articles. For every entity recognized by both methods, we created a vertex in our knowledge graph. We also applied additional manual 'filtering' of the resulting named entities. The procedure to fix the problems of the different spelling and some artifacts of NER and OIE that crowded the list of entities. Finding longer overlapping substrings with high frequencies we matched longer entities with their shorter "parents". The recognized vertexes were connected with an edge that had an estimate of sentiment and subjectivity calculated with TextBlob[4]. This naive approach yielded a hypergraph of named entities out of both data sources. The weights of the vertexes corresponded to the number of mentions of a given entity. The edges of the graph had three attributes: frequency, polarity, and subjectivity. To facilitate further research of news coverage and political discourse we share the gathered data[5].

3. Do You Know What I Know?

In this chapter, we explore the acquired knowledge graphs. In Section 3.1, we present a bird's-eye view of the graph, including key properties, and delve into the most contrasting entities and topics with varying coverage in two sources. Section 3.2 revisits the graphs, highlighting aspects crucial for differences in political discourse.

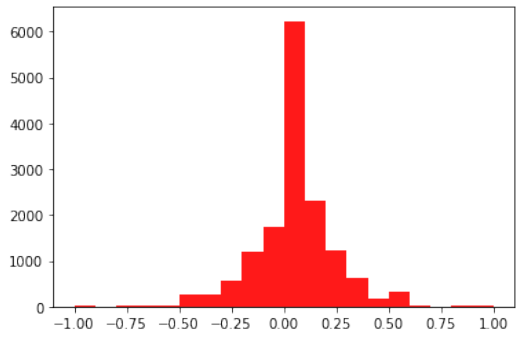


Figure 1: Breitbart News. Distribution of sentiment.

3.1. Bird's-eye View

Figures 5 – 6 show a visualization of two obtained graphs. One can see the divergence of topics:

[1] https://www.breitbart.com/
[2] https://www.nytimes.com/
[3] https://allennlp.org
[4] https://textblob.readthedocs.io/en
[5] https://shorturl.at/ntDOT

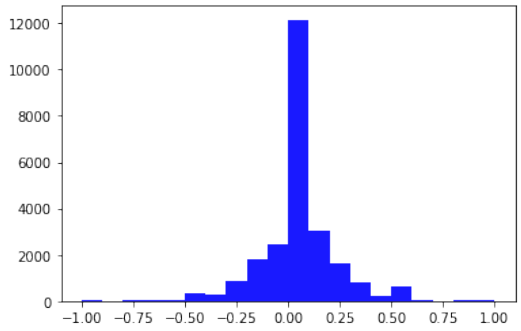


Figure 2: New York Times. Distribution of sentiment.

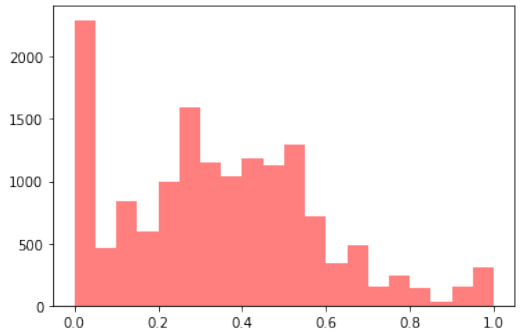


Figure 3: Breitbart News. Distribution of subjectivity.

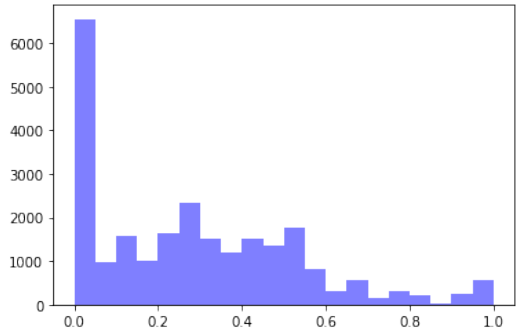


Figure 4: New York Times. Distribution of subjectivity.

Breitbart is more focused around certain personalities, while the New York Times extensively covers foreign affairs. Table 1 shows the first interesting and counter-intuitive result that one can draw when studying obtained graph representations: both media sources are "neutral" on average. Figures 1 – 2 show the distribution of polarity across all edges. The average neutral tone is not a consequence of negatively and positively charged news that balance each other. Distributions in Figures 3 – 4 do not only show that average sentiment across all edges is very close to zero for both graphs, but they also demonstrate and a vast majority of the analyzed relations are presented in a non-polarizing

Data	Radius	Diameter	Modularity
BN	6	11	0.43
NYT	7	13	0.53
	Average		
Data	**Path length**	**Polarity**	**Subjectivity**
BN	3.76	0.00	0.12
NYT	3.52	-0.00	0.08.

Table 1: Various parameters of the obtained graph representations. Both sources are neutral on average with Breitbart being just above and NYT just below zero average polarity. Breitbart tends to be more subjective, yet average subjectivity for both sources is at around 10%, with NYT a bit more objective.

way (at least to the extent to which modern NLP method can distinguish polarity). One can also see the corresponding distributions of subjectivity that are similar for both sources. For the NYT Spearman correlation between polarity and subjectivity is 31%, for Breitbart, it is 23%.

Both media sites try to present themselves to the reader as neutral on average and moderately subjective. This stands to reason: an average reader probably neither wants to feel that she wears rose-tinted glasses nor wants to constantly read that the doom is nigh. Majority of the news are neutral, extremely positive and extremely negative news are rare in both sources. At the same time both sources tend to point bias in the coverage "on the other side". Another interesting line of thought that could be developed when regarding Table 1 is the connection between right political actors and propagation of conspiracy theories, see, for example, (Hellinger, 2018). Indeed, the Breitbart graph has smaller modularity and comparable path length. This could imply a lower encapsulation of topics and a higher tendency to connect remote entities. Even a first bird's eye view gives several fundamental insights:

- when assessed formally both right and left media demonstrate qualitatively comparable behavior; they try to cover the news in a relatively neutral tone with a pinch of subjectivity;

- the coverage of various topics differs significantly; the entities that Brietbart constantly covers tend to be people and actors of domestic US politics, whereas NYT pays more attention to institutions and international affairs;

- the overall differences between formally obtained knowledge structures that could proxy right and left world-view are minute, despite our intuition telling us otherwise.

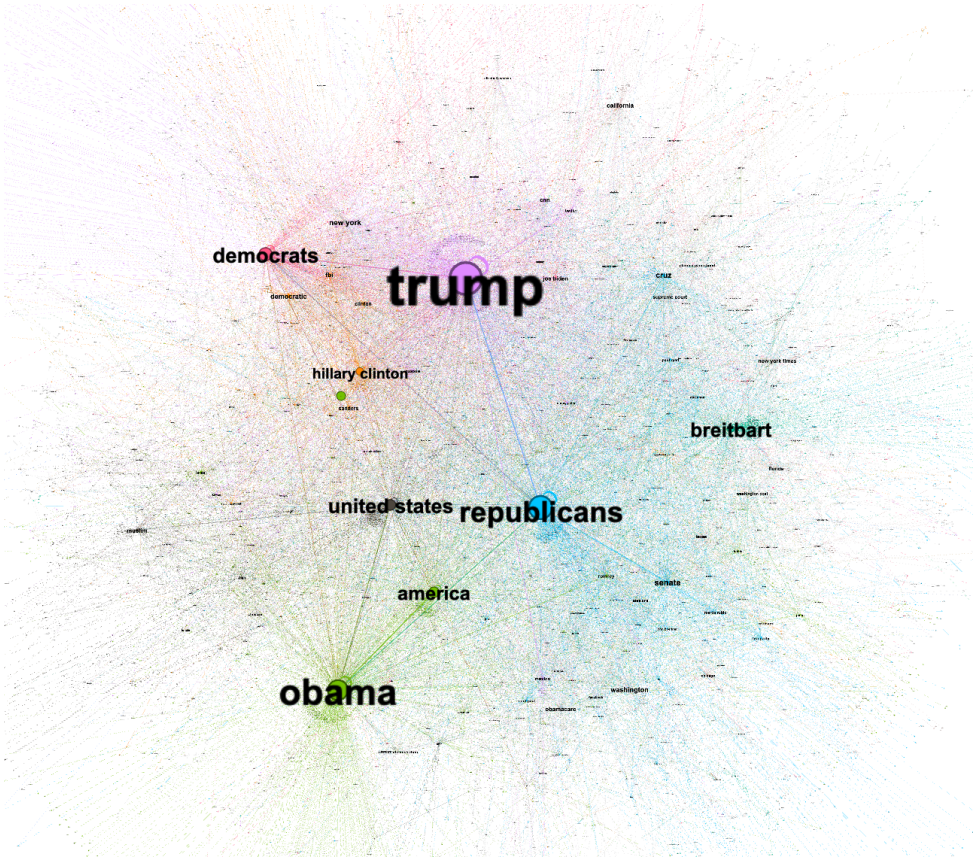


Figure 5: Breitbart News. Overall visualisation of two graphs extracted out of the media sources. The classes found with modularity analysis (Blondel et al., 2008) are highlighted with different colours. Breitbard has smaller number of classes and is centered around US political discourse.

3.2. Politics of Contrasts

Figure 7 shows a joint graph of the most polarized edges. These are the edges between entities for which the polarity in NYT and Breitbart has a different sign. Similarly, in Figure 8 one could see the most contrasting vertexes. These are the entities that have the highest average polarity of the adjacent edge. Effectively these are the representation of the polarizing topics and are covered with different polarity in both news sources.

An interesting difference between the graph of contrasting edges and the graph of contrasting nodes is that the former is mostly populated with domestic political actors, whereas the latter up to a large extent consists of entities connected with foreign affairs. This is interesting. Certain relationships between entities tend to be more polarizing for domestic issues and local politicians, yet when averaged over several such relationships across time the foreign affairs and institutions come forward. This is the same pattern that we saw earlier. One could speculate that contrasting edges highlight certain local events centered around specific politicians. Such events could be highly polarizing yet temporal. At the same time institutions and global affairs might not be as polarizing as a local scandal, yet the position of both sides on them is persistent, so when averaging across adjacent edges one sees Figure 8.

This highlights the fundamental difference between the sources. Though on macro-level both outlets prefer to stick to neutral coverage and refrain from subjectivity when it comes to certain entities and topics they provide different evaluations and tend to be more subjective in these cases. The combination of these two factors is extremely unfortunate since it facilitates social conflict. Indeed, every reader is perfectly convinced that her news source is relevant, objective, and non-biased. This also happens to be true in the vast majority of cases. Yet on a handful of key issues, the media takes a more polarizing and subjective position. Moreover, the local polarizing issues tend to be associated with personalities, while longer, fundamental differences are associated with institutions. This could be attributed to the idea of core political beliefs that could be less polarizing yet may be harder to change in the long run.

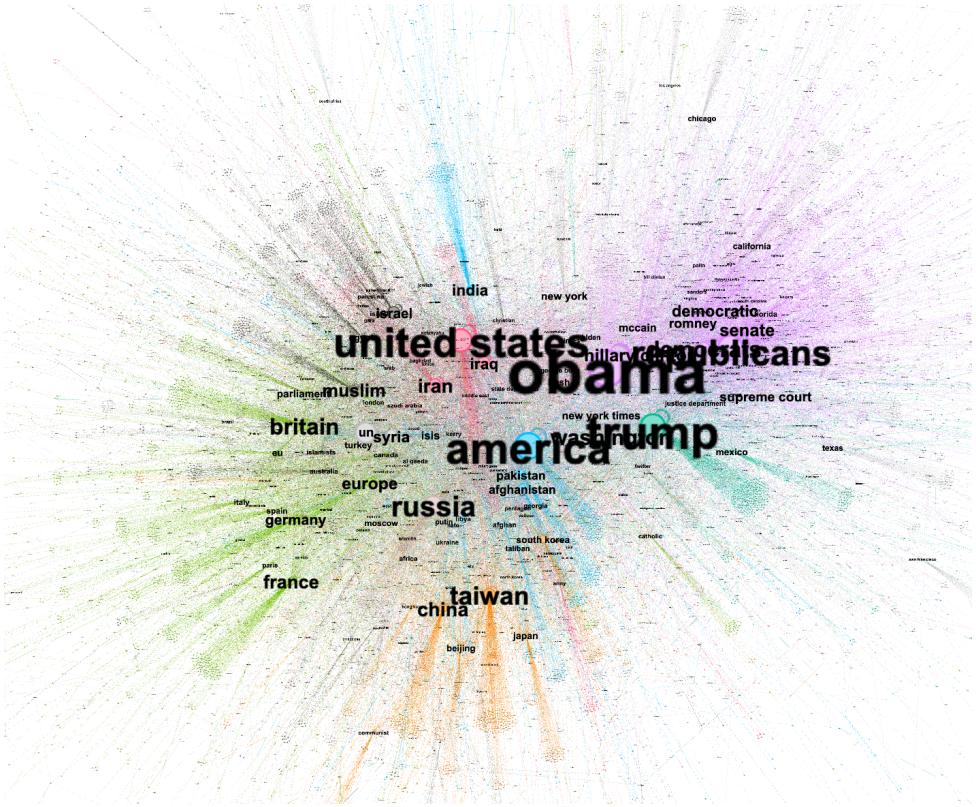


Figure 6: New York Times.Overall visualisation of two graphs extracted out of the media sources. The classes found with modularity analysis (Blondel et al., 2008) are highlighted with different colours. NYT graph has almost twice as many modularity classes and pays way more attention to foreign affairs. This could be partially attributed to the bigger size of the resulting graph, since NYT had more articles published in the studied time period.

4. Discussion

One of the key contributions of this paper is an attempt to demonstrate that an echo-chamber is not exactly a phenomenon based solely on the topology of human social networks. Using modern language processing methods and straight-forward knowledge representation we show that two different media sources paint two different pictures of the political reality. Yet these differences are less obvious than we tend to think and are more subtle. Surprisingly low average polarity and subjectivity for both knowledge structures are extremely intriguing. Assuming there is no ill will on the side of the publisher one can try to explain why the overwhelming amount of news articles try to be non-polarizing and non-subjective and with these attempts reinforce the echo-chambers around them without even trying. Could echo-chambers be a consequence of human psychological trust mechanisms on top of certain social structure formation? In (Levine, 2014) and (Clare and Levine, 2019) the authors discuss the truth-default theory. They demonstrate that when people cognitively process the content of others' communication, they typically do so in a manner characterized by unquestioned, passive acceptance. We could speculate that such behavior naturally transfers to the news sources. High neutrality and low subjectivity reinforce this truth-default. Since the preferred news outlet is often objective and neutral the reader tends to ignore or accept rare polarizing and subjective articles and dismiss the counter-argument of the other side, since in an overwhelming majority of the cases the criticism was not applicable. This might be wild speculation that demands further experimental verification. However, the very idea that echo-chamber formation could be attributed to the personal rather than collective behavior is new to our knowledge.

5. Conclusion

In this paper, we present the graphs of entities that correspond to two major "alt-right" and "liberal" news media and their coverage of the mentioned entities and relations between them. The graphs are obtained without any expert knowledge solely with NLP instruments and methods of knowledge representation. Analyzing obtained graphs we show that despite common intuition they exhibit a lot of structural similarities. We also highlight

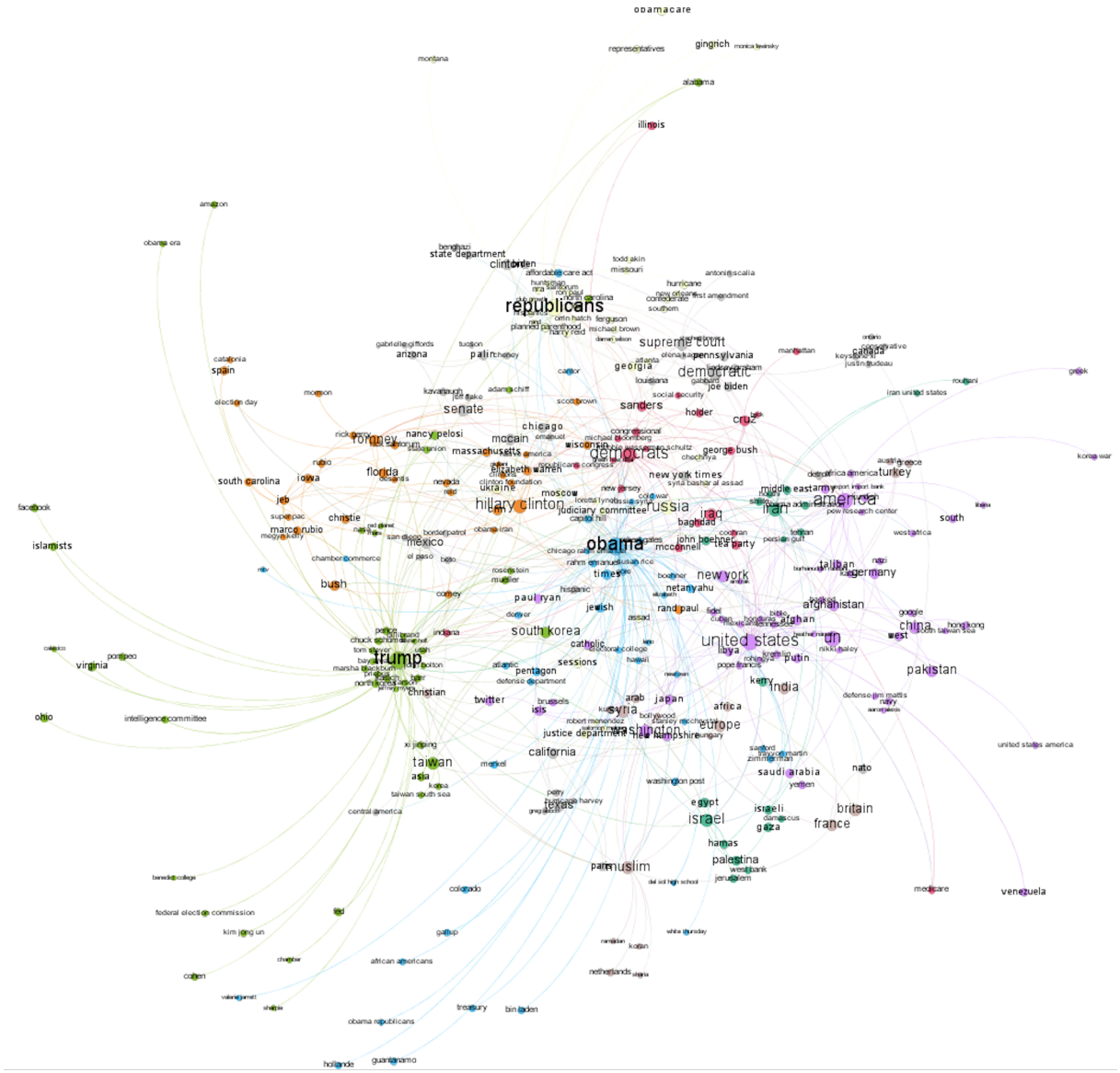


Figure 7: Sub-graph of contrasting edges. These are the edges for which the sign of polarity for BN and NYT is different.

fundamental differences that could be attributed to the formation of echo-chambers and certain biases on the world perception. We suggest that the formation of echo-chambers has more to do with the structure of information consumption and certain core beliefs of the individual rather than social structure that encompasses the aforementioned person.

Limitations

The study covers the period from 2008 to the Fall of 2019, excluding updates beyond 2019. It refrains from a detailed examination of the political aspects and perspectives of Breitbart News and New York Times readers, and it does not develop additional discussions on the global order. Considering recent global crises like wars, economic downturns in specific nations, and the worldwide impact of the COVID-19 pandemic, we anticipate that applying our methodology to recent-year data may produce slightly different findings. Nonetheless, in an effort to encourage transparent research in knowledge representation for social sciences, we provide access to our collected datasets.

Ethics Statement

Our work prioritizes transparency and relies on data collected from open sources. We refrain from making political judgments in our discussion notes to prevent discrimination and minimize potential societal harm.

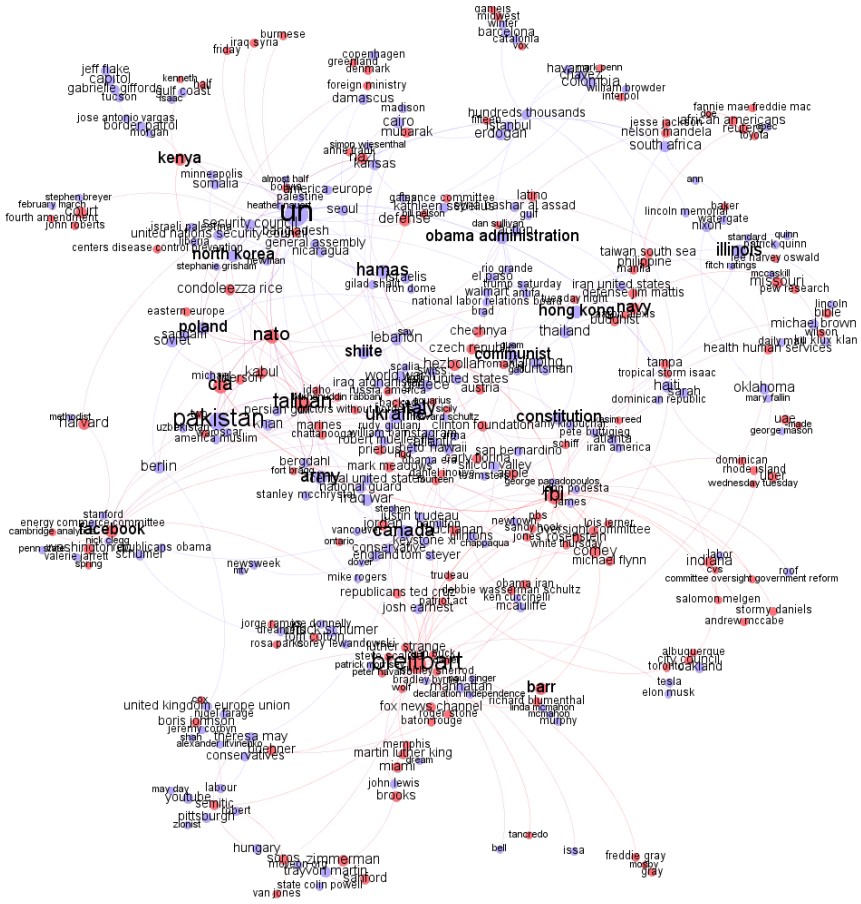


Figure 8: Sub-graph of contrasting vertexes. These are the vertexes for which the average of polarity of the adjacent edges is the highest. Blue nodes are shifted towards NYT, red — towards BN.

6. Bibliographical References

Al-Tawil M. Aljarah I. Faris H. Wongthongtham P. Chan K. Y. Abu-Salih, B. and A. Beheshti. 2021. Relational learning analysis of social politics using knowledge graph embedding. *Data Mining and Knowledge Discovery*, 35(1):1497–1536.

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Monica Anderson and Brooke Auxier. 2020. 55In *Pew Research Center*.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L_1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Sven Banisch and Eckehard Olbrich. 2019. Opinion polarization by learning from social feedback. *The Journal of Mathematical Sociology*, 43(2):76–103.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of*

the Association for Computational Linguistics, 5:135–146.

E. Borel. 1921. La theorie du jeu et les equations integrales a noyau symetrique. *Comptes rendus hebdomadaires des seances de l'Academie des sciences*, (173):1304–1308.

Romain Campigotto, Patricia Conde-Céspedes, and Jean-Loup Guillaume. 2014. A generalized and adaptive method for community detection.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Wei Chen, Xiao Zhang, Tengjiao Wang, Bishan Yang, and Yi Li. 2017. Opinion-aware knowledge graph for political ideology detection. In *IJCAI*, pages 3647–3653.

P. Chilton. 2004. *Analysing political discourse: Theory and practice*. Routledge.

David D Clare and Timothy R Levine. 2019. Documenting the truth-default: The low frequency of spontaneous unprompted veracity assessments in deception detection. *Human Communication Research*, 45(3):286–308.

Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. 2014. Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of communication*, 64(2):317–332.

Rosaria Conte, Nigel Gilbert, Giulia Bonelli, Claudio Cioffi-Revilla, Guillaume Deffuant, Janos Kertesz, Vittorio Loreto, Suzy Moat, J-P Nadal, Anxo Sanchez, et al. 2012. Manifesto of computational social science. *The European Physical Journal Special Topics*, 214(1):325–346.

James W. Cooley and John W. Tukey. 1965. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301.

Pieter Delobelle, Murilo Cunha, Eric Cano, Jeroen Peperkamp, and Bettina Berendt. 2019. Computational ad hominem detection. pages 203–209.

Felix Gaisbauer, Armin Pournaki, Sven Banisch, and Eckehard Olbrich. 2023. Grounding force-directed network layouts with latent space models. *Journal of Computational Social Science*, pages 1–33.

Kiran Garimella, Gianmarco Morales, Aristides Gionis, and Michael Mathioudakis. 2016. Quantifying controversy in social media. pages 33–42.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2017. Cross-lingual classification of topics in political texts. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 42–46.

David Goldie, Matthew Linick, Huriya Jabbar, and Christopher Lubienski. 2014. Using bibliometric and social media analyses to explore the "echo chamber" hypothesis. *Educational Policy*, 28(2):281–305.

T. Graham, D. Jackson, and M. Broersma. 2016. New platform, old habits? candidates use of twitter during the 2010 british and dutch general election campaigns. *New Media & Society*, 18(5):765–783.

O. Gross and R. Wagner. 1950. A continuous colonel blotto game. RAND Research Memorandum.

Fangjian Guo, Charles Blundell, Hanna Wallach, and Katherine Heller. 2015. The bayesian echo chamber: Modeling social influence via linguistic accommodation. In *Artificial Intelligence and Statistics*, pages 315–323.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Lisa Harris and Paul Harrigan. 2015. Social media in politics: The ultimate voter engagement tool or simply an echo chamber? *Journal of Political Marketing*, 14(3):251–283.

Daniel C Hellinger. 2018. *Conspiracies and conspiracy theories in the age of trump*. Springer.

Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PloS one*, 9:e98679.

Julie Jiang, Xiang Ren, and Emilio Ferrara. 2021. Social media polarization and echo chambers: A case study of covid-19.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. 2022. Logical fallacy detection. page 1.

Kristen Johnson and Dan Goldwasser. 2018. Classification of moral foundations in microblog political discourse. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 720–730.

A. Jungherr. 2014. The logic of political coverage on twitter: Temporal dynamics and content. *Journal of Communication*, 64(2):239–259.

Jonathan P Kastellec and Eduardo L Leoni. 2007. Using graphs instead of tables in political science. *Perspectives on politics*, pages 755–771.

Y. Kim. 2014. Convolutional neural networks for sentence classification. ArXiv:1408.5882.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

D. Kovenock and B. Roberson. 2015. Generalizations of the general lotto and colonel blotto games. CESifo Working Paper 5291.

Werner Krause, Pola Lehmann, Jirka Lewandowski, Theres Matthieß, Nicolas Merz, Sven Regel, and Annika Werner. 2018. Manifesto corpus. *WZB Berlin Social Science Center*.

Onawa P Lacewell and Annika Werner. 2013. Coder training: Key to enhancing coding reliability and estimate validity.

Man Lan, Chew Lim Tan, Jian Su, and Yue Lu. 2008. Supervised and traditional term weighting methods for automatic text categorization. *IEEE transactions on pattern analysis and machine intelligence*, 31(4):721–735.

J.-F. Laslier and N. Picard. 2002. Distributive politics and electoral competition. *Journal of Economic Theory*, (103):106–130.

David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

P. Lehmann, T. Matthieß, N. Merz, S. Regel, and A. Werner. 2017. Manifesto corpus 2017-1. WZB Berlin Social Science Center.

Timothy R Levine. 2014. Truth-default theory (tdt) a theory of human deception and deception detection. *Journal of Language and Social Psychology*, 33(4):378–392.

Steven Loria et al. 2018. textblob documentation. *Release 0.15*, 2(8):269.

Shawn Martin, W. Brown, Richard Klavans, and Kevin Boyack. 2011. Openord: An open-source toolbox for large graph layout. *Proc SPIE*, 7868:786806.

Nicolas Merz, Sven Regel, and Jirka Lewandowski. 2016. The manifesto corpus: A new resource for research on political parties and quantitative text analysis. *Research & Politics*, 3(2):2053168016643346.

Slava Mikhaylov, Michael Laver, and Kenneth R Benoit. 2012. Coder reliability and misclassification in the human coding of party manifestos. *Political Analysis*, 20(1):78–91.

R.B. Myerson. 1993. Incentives to cultivate minorities under alternative electoral systems. *American Political Science Review*, 87:856–869.

Usman Naseem, Imran Razzak, Katarzyna Musial, and Muhammad Imran. 2020. Transformer based deep intelligent contextual embedding for twitter sentiment analysis. *Future Generation Computer Systems*, 113.

R. Neuman, L. Guggenheim, S. Mo Jang, and S.Y. Bae. 2014. The dynamics of public attention: Agenda setting theory meets big data. *Journal of Communication*, 64(2):193–214.

Sri Nurdiati and Cornelis Hoede. 2008. 25 years development of knowledge graph theory: the results and the challenge. *Memorandum*, 1876(2):1–10.

A. Osorio. 2013. The lottery blotto game. *Economics Letters*, 120(2):164–166.

I. Parker. 2014. *Discourse Dynamics (Psychology Revivals): Critical Analysis for Social and Individual Psychology*. Routledge.

Tiago P Peixoto. 2020. Latent poisson models for networks with heterogeneous density. *Physical Review E*, 102(1):012309.

Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. *ArXiv*, abs/1705.00108.

John Petit, Cong Li, and Khudejah Ali. 2020. Fewer people, more flames: How pre-existing beliefs and volume of negative comments impact online news readers' verbal aggression. *Telematics and Informatics*, 56:101471.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Arsenii Rasov, Ilya Obabkov, Eckehard Olbrich, and Ivan P Yamshchikov. 2020. Text classification for monolingual political manifestos with words out of vocabulary. In *COMPLEXIS*, pages 149–154.

B. Roberson. 2006a. The colonel blotto game. *Economic Theory*, 29(1):1–24.

B. Roberson. 2006b. Pork-barrel politics, discriminatory policies and fiscal federalism. Social Science Research Center Berlin (WZB).

Richard Rogers. 2013. *Digital methods*. MIT press.

Wilbert Samuel Rossi, Jan Polderman, and Paolo Frasca. 2018. The closed loop between opinion formation and personalised recommendations.

Martin Rosvall and Carl T Bergstrom. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the national academy of sciences*, 105(4):1118–1123.

Dhavan V Shah, Joseph N Cappella, and W Russell Neuman. 2015. Big data, digital media, and computational social science: Possibilities and perils. *The ANNALS of the American Academy of Political and Social Science*, 659(1):6–13.

Jesse Shore, Jiye Baek, and Chrysanthos Dellarocas. 2018. Network structure and patterns of information diversity on twitter. *MIS Quarterly*, 42.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *NAACL-HLT*.

Shivashankar Subramanian, Trevor Cohn, and Timothy Baldwin. 2018. Hierarchical structured model for fine-to-coarse manifesto text analysis. 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 1964–1974.

Shivashankar Subramanian, Trevor Cohn, Timothy Baldwin, and Julian Brooke. 2017. Joint sentence-document model for manifesto text analysis. In *Proceedings of the Australasian Language Technology Association Workshop 2017*, pages 25–33.

Andon Tchechmedjiev, Pavlos Fafalios, Katarina Boland, Malo Gasquet, Matthäus Zloch, Benjamin Zapilko, Stefan Dietze, and Konstantin Todorov. 2019. Claimskg: a knowledge graph of fact-checked claims. In *International Semantic Web Conference*, pages 309–324. Springer.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. *arXiv preprint cs/0607062*.

Richard L Tweedie, Kerrie L Mengersen, and John A Eccleston. 1994. Garbage in, garbage out: can statisticians quantify the effects of poor data? *Chance*, 7(2):20–27.

Suzan Verberne, Eva D'hondt, Antal van den Bosch, and Maarten Marx. 2014. Automatic thematic classification of election manifestos. *Information Processing & Management*, 50(4):554–567.

Zhouxia Wang, Tianshui Chen, Jimmy Ren, Weihao Yu, Hui Cheng, and Liang Lin. 2018. Deep reasoning with knowledge graph for social relationship understanding. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 1021–1028.

Michael D Ward, Katherine Stovel, and Audrey Sacks. 2011. Network analysis and political science. *Annual Review of Political Science*, 14:245–264.

A. Washburn. 2013. Blotto politics. *Operations Research*, 61(3):532–543.

Ivan P Yamshchikov and Sharwin Rezagholi. 2019. Elephants, donkeys, and colonel blotto. In *COMPLEXIS*, pages 113–119.

Cäcilia Zirn, Goran Glavaš, Federico Nanni, Jason Eichorts, and Heiner Stuckenschmidt. 2016. Classifying topics and detecting topic shifts in political manifestos.

Analyzing Conflict Through Data: A Dataset on the Digital Framing of Sheikh Jarrah Evictions

Anatolii Shestakov, Wajdi Zaghouani

Hamad Bin Khalifa University

ansh33161@hbku.edu.qa, wzaghouani@hbku.edu.qa

Abstract

This study empirically investigates the role of social media in tracing the evolution of the May 2021 Israeli-Palestinian crisis, centered on the Sheikh Jarrah evictions. Analyzing a dataset of 370,747 English tweets from 120,173 users from May 9-21, 2021, the research employs a mixed-methods approach combining computational techniques and qualitative content analysis. Findings support the hypothesis that social media interactions reliably map crisis dynamics, as evidenced by hashtags like #SaveSheikhJarrah corresponding to critical shifts, though virality did not correlate with hashtag use. In contrast to prior sentiment-focused studies, the context-driven analysis reveals influencers and state actors shaping polarized narratives along geopolitical lines, with high-profile voices backing Palestinian solidarity while Israeli state accounts endorsed military operations. Evidence of a transcontinental cybercampaign emerged, albeit with limitations due to the English language scope and potential biases from data collection and keyword choices. The study contributes empirical insights into the mediatization of armed conflicts through social media's competing narratives and information flows within the Israeli-Palestinian context. Recommendations for future multilingual, multi-platform analyses are provided to address limitations.

Keywords: Sheikh Jarrah, Social Media, Cybercampaign, Content Analysis, Digital Framing

1. Introduction

The Sheikh Jarrah neighborhood, a predominantly Palestinian area in East Jerusalem, has been a focal point of property disputes following the Israeli occupation in 1967. In May 2021, tensions escalated when Israeli authorities threatened to evict Palestinian residents, triggering violent clashes. This incident was a flashpoint in the longstanding Israeli-Palestinian conflict, which has been extensively documented and debated across traditional and social media platforms. The proliferation of social media has provided previously marginalized voices an opportunity to participate in these discourses, shaping narratives and mobilizing support.

This study investigates the unfolding of the May 2021 escalation, known as the Gaza crisis, through an analysis of Twitter interactions. The intensity of user reactions prompted the formation of the #SaveSheikhJarrah movement, generating a substantial corpus of tweets during this period. By employing a mixed-methods approach combining computational techniques and qualitative analysis, this research aims to provide empirical insights into the dynamics of the 2021 crisis within the broader context of the Israeli-Palestinian confrontation since 1948. While existing literature has primarily focused on earlier events, this study offers a contemporary perspective informed by the analysis of user-generated content on Twitter.

The methodological framework encompasses the examination of influential accounts, information flows, linguistic patterns, and hashtag usage, facilitating a nuanced understanding of the discourse dynamics on this social media platform. By triangulating quantitative and qualitative findings, the study seeks to extend theoretical frameworks on information flows and narratives surrounding the Middle East conflict.

A central focus of this investigation is the polarizing nature of media narratives during the Israeli-Palestinian crisis, interweaving findings on cyber-activism within the context of armed conflicts in the region. Analyzing Twitter reactions poses challenges due to the extensive media coverage and potential manipulation of information flows.

To comprehensively examine this phenomenon, the following hypothesis is evaluated: **H1:** Social media interactions constitute a reliable source for measuring the development and dynamics of conflicts.

To address this hypothesis, the study explores the following research questions:

RQ1: What were the dynamics of user activity during the 2021 Israeli-Palestinian crisis, and how can qualitative and quantitative analyses shed light on user typologies and behaviors on this new media platform?

Proceedings of the Second Workshop on Natural Language Processing for Political Sciences (PoliticalNLP 2024), pages 55–67

RQ2: What sentiment and thematic patterns prevail in the collected tweets, and how can corpus analysis techniques based on linguistic patterns and keywords elucidate the predominant moods and styles of users?

RQ3: Is it possible to establish the geographical locations of cyber-activists during the Sheikh Jarrah eviction campaign by contextualizing and summarizing the accumulated data to ascertain spatial patterns of pro-Palestinian and pro-Israeli users?

RQ4: How effective are hashtag statistics in evaluating the stages of the conflict, and can analyses of hashtag usage data generate visual representations of the conflict's evolution?

This interdisciplinary work draws upon diverse fields, including comparative history, humanities, cultural studies, data analysis, and digital humanities, to depict the intricate phenomenon of contemporary cyber-activism. The fundamentals of postmodernist methodology, primarily the theories of Foucault and researchers of public communications (Ciszek, 2016; Holtzhausen, 2002; Holtzhausen & Voto, 2002; Kent, 2002), are described. The combination of computational research and critical close-reading approaches aims to contextualize findings and mitigate the tendency towards purely quantitative analyses in information studies (Felt, 2016). The paper is structured as follows: Section 2 provides a review of relevant literature, Section 3 outlines the methodology, Section 4 presents the results, Section 5 discusses limitations and future directions, and Section 6 concludes the study.

2. Related Work

The Israeli occupation of Palestine has been a subject of extensive research, with scholars employing diverse theoretical frameworks and methodologies to analyze its complex dynamics. Postmodernist approaches, which reject the notion of a single objective truth and emphasize the role of power relations in shaping discourse (Deetz, 2001), have been particularly influential in this context. By focusing on marginalized groups' resistance to power structures and their ability to shape narratives (Holtzhausen, 2002), postmodernist perspectives offer valuable insights into the distribution of knowledge and the interplay between power and resistance in the Israeli occupation of Palestine.

The advent of social media has profoundly impacted how conflicts are perceived, experienced, and contested. Echo chambers, where users are exposed primarily to opinions that align with their own beliefs (Garimella et al., 2018), have emerged as a significant phenomenon, challenging the ideal of free information circulation (An et al., 2014; Cacciatore et al., 2016; Grömping, 2014; Lawernce et al., 2010). Researchers have demonstrated the influence of social media on real-life political polarization, which is often intensified by ongoing developments (Azzimonti and Fernandes, 2018; Du and Gregory, 2016). However, previous studies on Gaza cybercampaigns have not specifically addressed the polarization between pro-Israeli and pro-Palestinian camps (Mtchedlidze, 2019), a gap that this study aims to fill by examining the distinct polarization between opposing factions.

In times of crisis, the spread of misinformation, whether intentional or unintentional, can exacerbate the risk of individuals being influenced by conspiracy theories, rumors, and falsehoods (Bodaghi and Oliveira, 2020; Lewandowsky et al., 2013). The COVID-19 pandemic has brought renewed attention to the study of misinformation, particularly in the context of public health policies (McGlynn et al., 2013; Naeem and Ozuem, 2021; Shahi et al., 2021). However, misinformation has also been examined in earlier crises, such as the 2013 Boston Marathon bombings (Huang et al., 2015) and hurricanes Harvey and Irma (Hunt et al., 2020), as well as in socio-political events like elections (Kušen and Strembeck, 2018; Sanderson et al., 2021). These studies have employed large social media datasets and techniques such as topic modeling (Jamison et al., 2020) and deep learning (Ajao et al., 2018) to identify misinformation patterns and understand human behavior in crisis situations. Social media data has been increasingly used to analyze real-world events, including Israeli occupation of Palestine. Sarraj et al. (2016) observed a spike in Twitter activity during the 2014 Israel-Palestine war, while Zeitzoff et al. (2015) explored the connections between online interactions and offline violence in the context of Iran-Israel relations. Zeitzoff (2011) conducted a macrodynamic analysis of Twitter activity during the 2008-2009 Gaza conflict, identifying two distinct peaks in user reactions that corresponded to critical moments in the conflict. The rise of ISIS, the occupation of Crimea, and the Trump elections also saw a dramatic increase in the speed of information flow on social media platforms (Zeitzoff, 2018). Zeitzoff (2017) noted that the Israeli Defence Forces (IDF) adapted their campaigns based on hashtag activity on social networks, highlighting the strategic importance of social media in modern conflicts. The role of digital media in the Middle East and North Africa (MENA) region has been extensively studied, particularly in the context

of the Arab Spring and other regional uprisings. Twitter has been credited with playing a significant role in these events, facilitating the expression of dissent and mobilization efforts. In this context, Israel has sought to challenge neighboring Arab states by establishing a strong presence on social media and promoting user-generated content that advances its position (Stein, 2012). Studies have examined the impact of Israeli state-run social media accounts on public opinion regarding the Gaza war (Seo, 2014), Israeli digital diplomacy (Manor and Crilley, 2018), and war legitimation strategies (Simonsen, 2019). In contrast, the online pro-Palestinian community has been characterized by more decentralized and unconstrained user activity.

The integration of media into warfare has evolved over time, with the Gulf War marking the rise of telecommunications (stage one) and the conflicts in Ukraine and Libya exemplifying the decentralization of internet media (stage two). Hoskins and O'Loughlin (2015) theorized that states would exert greater control over information in the third "arrested war" stage, a trend that could be observed in the Gaza conflict due to Israel's stronger information appropriation. While pro-Palestinian media remains in stage two, Israel has moved towards stage three with its "arrested" state pages, which Stein (2014) described as "militarised social media" or "digital militarism" due to their attempts at projecting a sense of "everydayness" (Hoskins and O'Loughlin, 2015). Personal perspectives of wartime have also been explored through social media. Martínez García (2017) analyzed the power of a digital diary written by a seven-year-old Syrian girl on Twitter, highlighting the importance of documentary activism during conflict. Tawil-Souri and Aouragh (2014) examined Israeli-Palestinian digital disputes, questioning the extent to which online activism translates into physical resistance. Their study contextualized Palestinian online resistance within the broader framework of digital anticolonial discourse, noting that while the internet and social media have provided new tools for resistance, they have also given rise to "new forms of colonialism" (Tawil-Souri and Aouragh, 2014).

The role of gatekeeping in information dissemination has undergone significant changes with the advent of the internet. While journalists traditionally served as the primary gatekeepers (Wallace, 2018), the rise of social media has shifted this power to corporate platforms (Wallace, 2018; Kent, 2014). Although citizens now have the potential to act as gatekeepers, they are still subject to the algorithms and policies of these platforms (Wallace, 2018). The silencing of social media platforms by regimes can lead to the spread of misinformation, as observed during the Arab Spring, Iranian Green Movement, and Syrian protests, when authorities resorted to censorship and internet access blocking (Fekete and Warf, 2013; Golkar, 2011; Shehabat, 2012).

In the case of Israeli occupation of Palestine in 2021, the control over media channels was less overt but still significant, taking the form of content moderation by tech giants. Alimardani and Elswah (2021) discussed this phenomenon, termed "digital orientalism," in the context of cyber-activism in the MENA region. They highlighted issues such as the deletion of Syrian war crimes reports on YouTube, the erasure of regional differences through automated Arabic translations, unequal access to ad data for Tunisian activists on Facebook, and the lack of regional offices for Arab states. These factors contribute to the threat of digital discrimination on social media. Given the evidence of linguistic bias against Palestinians in traditional media reporting (Barkho and Richardson, 2010), decentralized media has gained importance as a means of rebalancing the narrative (Shreim and Dawes, 2015).

Recent studies have employed computational methods and qualitative analysis to investigate the dynamics of social media interactions during the Israeli occupation of Palestine. Alam et al. (2021) modeled the perspectives of various stakeholders in combating the COVID-19 infodemic, emphasizing the importance of considering multiple viewpoints when analyzing crisis-related discourse on social media. Hasanain et al. (2023) focused on detecting persuasion techniques and disinformation in Arabic text, underscoring the need for robust computational methods to identify manipulative content in social media discussions.

The CLEF-2018 CheckThat! Lab (Atanasova et al., 2018) and the CLEF-2023 CheckThat! Lab (Alam et al., 2023) have explored the automatic identification and verification of political claims, as well as the assessment of check-worthiness in multimodal and multigenre content, providing valuable methodological insights for analyzing political framing and misinformation in social media data. Several studies have specifically examined political framing in the context of crises and conflicts. Shurafa et al. (2020) investigated the US COVID-19 blame game on social media, demonstrating how computational techniques can be employed to

uncover patterns of political framing. Laabar and Zaghouani (2024) created an annotated dataset of stance, sentiment, and emotion in Facebook comments related to Tunisia's political measures, showcasing the value of multi-dimensional analysis in understanding public opinion on social media.

The creation and analysis of large-scale, multi-dialect Arabic social media corpora have also been a focus of recent research. Zaghouani and Charfi (2018) presented the Arap-tweet corpus, which includes gender, age, and language variety identification, while Alam et al. (2021) and Shaar et al. (2021) developed datasets and shared tasks for detecting COVID-19 misinformation and censorship in Arabic social media content. Author profiling, deception detection, and irony detection in Arabic social media have also received attention, as evidenced by the survey conducted by Rosso et al. (2018) and the shared task organized by Rangel et al. (2019). These studies highlight the importance of considering linguistic and cultural nuances when analyzing Arabic social media content.

Building upon these previous works, our study aims to contribute to the growing body of research on social media discourse analysis in the context of the Israeli occupation of Palestine. By leveraging computational methods and qualitative analysis, we seek to uncover patterns of political framing, misinformation, and polarization in the Twitter discourse surrounding the May 2021 crisis, while also considering the unique challenges and opportunities presented by the Arabic language and the region's socio-political context. We will examine features of modern digitalized activism in the #SaveSheikhJarrah campaign, which challenges social and political inequalities and adapts to a new media paradigm in the face of limited mediation in the region. By applying computational tools, this study analyzes information flow, compares it with existing theories, and operationalizes big data and user-generated content within media trends.

3. Methodology

This study employs a multi-faceted approach to analyze publicly accessible social media content related to the May 2021 Israeli-Palestinian conflict, while adhering to user privacy protocols. The research methodology consists of three primary phases: data collection, quantitative analysis, and qualitative close reading. This approach enables a comprehensive examination of the data's connotative meanings, particularly at the lexical level, which is of significant interest to scholars in the field of Digital Humanities.

3.1 Data Collection

To ensure the relevance of the collected tweets to the May 2021 confrontation, a precise search query was established, focusing on the period from May 9 to May 21, 2021, which encompassed the conflict's preliminary and terminal phases. Using Python for automation, tweets were harvested based on the keywords "Gaza," "Israel," and "Palestine," which were central to the discussions during the escalated conflict. The resulting dataset comprised 370,747 tweets from 120,173 distinct users, with retweet duplicates excluded to maintain data integrity and relevance. The dataset can be obtained by contacting the authors[1].

The corpus predominantly features tweets in Latin script, deliberately excluding Arabic and Hebrew to mitigate language bias and represent international discourse authentically. The data was collected in September 2021 and converted to UTF-8 format for accessibility and further analysis. It is important to note that the dataset may not be exhaustive, as some tweets may have been deleted or suspended after the collection phase. Before proceeding with the analysis, the corpus was filtered to remove URL links and user mentions (denoted by "@") to refine the dataset for detailed examination.

3.2 Data Analysis

The study employs an integrated methodological approach, combining quantitative and qualitative strategies to analyze and interpret the social media data, yielding a multifaceted descriptive analysis. Advanced computational tools and algorithms are selectively applied at various stages of the analysis to meet the research's specific needs. Identifying key influencers within the discourse was a crucial aspect of the analysis. A Python-automated application programming interface (API) was used to quantify the number of followers, determining the influential power of various accounts. Manual examination provided deeper insights into the nature of these accounts, classifying them into categories such as state-run entities, journalists, personal pages, and others to elucidate their roles in the discourse. To understand the sentiments and thematic directions of the discussions, a detailed contextual analysis of the tweets was

[1] To request the dataset for research purposes, please fill the following form: https://forms.gle/S9fZtYjAyLAqFsH19

conducted, assessing the polarity and conflict perspectives embodied in the social media interactions. Voyant Tools, a corpus linguistics tool, was employed to extract and visualize keyword trends, facilitating a nuanced examination of the lexical patterns and thematic occurrences within the data. This tool was also instrumental in mapping the distribution of hashtags within a curated subset of the data, focusing on the semantics and usage patterns of these hashtags.

The analysis focused on an English tweets corpus to encapsulate the global perspective and linguistic nuances, providing insights into the international discourse. The AntConc tool, a corpus analysis utility, was used to perform a randomized exploration of keywords in context (KWIC), enabling a meticulous close reading of the text fragments. The analytical process was enhanced by adopting Braun and Clarke's (2006) thematic analysis framework, a qualitative content analysis method that facilitated a deeper understanding of the underlying themes within the data.

The synthesis of these quantitative and qualitative methodologies aimed to construct a comprehensive and multifaceted interpretation of the user-generated content pertaining to the events of the May confrontation. This holistic approach to data analysis ensures a robust and insightful exploration of the digital discourse, providing valuable perspectives on the dynamics and sentiments expressed during the conflict.

4. Results

4.1 4.1 Information Diffusion and Gatekeeping

To investigate information diffusion trends during the conflict, top accounts within the discussion were analyzed based on the number of followers they attracted and engaged with (see Table 1). The top 10 accounts by follower rate consist predominantly of media giants such as CNN and BBC, with an exception in line 6, where British artist Zayn Malik is positioned. While news accounts depicted information neutrally, Malik stood forward to support Palestine, placing him among the most influential accounts with a distinct position regarding the conflict. Interestingly, the most engaged tweets were found to be from accounts with less than 50,000 followers, with some accounts having less than 500 followers reaching broad audiences with retweet numbers up to 46,000 times. This finding suggests that the network empowered users with both highly developed information channels and relatively humble follower numbers.

Account name	Username	Followers[2]	Type
CNN Breaking News	@cnnbrk	61,349,476	media
CNN	@CNN	54,608,606	media
The New York Times	@nytimes	50,465,046	media
BBC Breaking News	@BBCBreaking	48,173,456	media
BBC News (World)	@BBCWorld	32,856,126	media
Zayn	@zaynmalik	30,976,562	artist[3]
The Economist	@TheEconomist	25,807,028	media
Reuters	@Reuters	23,760,876	media
Fox News	@FoxNews	20,202,832	media
CNN_EN Español	@CNNEE	20,139,173	media

Table 1: Identified Twitter accounts with the highest number of followers

Conflict side	Username	Retweets[4]	Type
Pro-Palestinian	@zaynmalik	190,106	artist
Pro-Palestinian	@godsxm	81,171	personal page
Pro-Palestinian	@godsxm	79,091	personal page
Pro-Palestinian	@JatIkhwan	68,132	personal page
Pro-Palestinian	@HausofHilton	59,972	personal page
Pro-Palestinian	@Mahrez22	58,500	sportsman
Pro-Palestinian	@velvetbiased	54,859	personal page
Pro-Palestinian	@WAYSTARFILMS	49,886	no longer exists
Pro-Palestinian	@MiddleEastEye	49,525	media
Pro-Palestinian	@EoinHiggins_	47,921	journalist

Table 2: Frequency of accounts with the most retweeted tweet

[2] Follower numbers are correct as of the dataset creation date.

[3] The only page on the list with a defined position about the conflict (pro-Palestinian).

[4] Retweets number at the moment of the dataset creation

4.2 Accounts Analysis

The most prominent accounts were further analyzed using different criteria, such as retweet count (Table 2), quote count (Table 3), like count (Table 4), and reply count (Table 5). The analysis revealed that accounts expressing political opinions and personal judgments tended to have higher retweet counts, with the most retweeted accounts pushing pro-Palestinian support. In terms of quote count, user behavior changed, with only one page supporting the pro-Palestinian side while the remaining accounts took the pro-Israeli position (Table 3). The Israeli state-run account occupied three positions in the table, while Israeli MFA-related accounts placed fifth and sixth.

Conflict side	Username	Quotes	Type
Pro-Israeli	@Israel	53,256	state-run page
Pro-Israeli	@AndrewYang	36,933	politician
Neutral	@GretaThunberg	36,369	activist
Pro-Israeli	@Mike_Pence	31,690	politician
Pro-Israeli	@IdoDaniel	30,118	media-expert
Pro-Israeli	@IdoDaniel	23,216	media-expert
Pro-Israeli	@Israel	22,587	state-run page
Pro-Israeli	@WhiteHouse	21,375	state-run page
Pro-Palestinian	@frhamlna	17,385	personal page
Pro-Israeli	@Israel	13,562	state-run page

Table 3: Accounts with the highest quoted tweet count

Conflict side	Username	Likes	Type
Pro-Palestinian	@zaynmalik	659,858	artist
Pro-Palestinian	@Mahrez22	192,414	sportsman
Pro-Palestinian	@BernieSanders	167,449	politician
Pro-Palestinian	@AOC	145,458	state representative
Pro-Palestinian	@zalx_	139,364	personal page
Pro-Israeli	@Israel	122,491	state-run page
Pro-Palestinian	@ElNennY	121,464	sportsman
Pro-Palestinian	@MiddleEastEye	114,794	media
Pro-Palestinian	@EoinHiggins_	107,871	journalist
Pro-Palestinian	@thisisnotmaha	106,299	personal page

Table 4: Accounts with the highest like count

Analysis of tweets by likability (Table 4) demonstrated pro-Palestinian moods among the most-liked tweets, with a diverse user demography, including artists, sportsmen, and personal accounts. The official Israeli state-run page also reached a broad user audience, positioning sixth among predominantly pro-Palestinian accounts.

The accounts filtered by reply count (Table 5) presented the most diverse group of accounts, with people from various occupations represented.

Conflict side	Username	Replies	Type
Pro-Israeli	@Mike_Pence	64,361	politician
Pro-Israeli	@Israel	37,091	state-run page
Pro-Israeli	@IdoDaniel	36,294	media-expert
Pro-Israeli	@AndrewYang	35,966	politician
Pro-Palestinian	@zaynmalik	35,863	artist
Pro-Israeli	@WhiteHouse	23,017	state-run page
Pro-Palestinian	@jeremycorbyn	22,536	politician
Pro-Palestinian	@djsnake	20,280	artist
Pro-Israeli	@Israel	16,557	state-run page
Pro-Israeli	@Israel	15,769	state-run page

Table 5: Accounts with the highest reply count.

4.3 Geographical Analysis

Regarding users' geographic locations, 1,418 unique users with automatically set geolocation (around 1%) were identified. Although this data is credible, it is not representative. 33% of users self-identified their locations (more than 39,000 places), even though the data often contained imaginary places or only emojis. After filtering this data, the map in Figure 1 can be considered, confirming the broad inclusion of the #SaveSheikhJarrah campaign covering most countries worldwide. However, these results should be considered strictly illustrative rather than precise, as noted by Voyant Tools developers (Sinclair and Rockwell, 2016).

4.4 Short-term Dynamics and Language Analysis

Although this study does not provide hourly visual data, an illustration of how user activity changed according to real-life events and

militant processes during the 11 days was created (figure not included due to space restrictions). The graph, divided into day and night parts, shows that Israeli-Palestinian discourse attracted people from around the globe, ensuring comparatively equal activity during the 24 hours per day.
Corpus data was collected using Python automation, scraping all tweets containing keywords or hashtags. Although the study targeted English content, 28% of the content was published in languages other than English.

Figure 1: The Map of User Activity (Based on Self-Indicated Locations)

4.5 Frequently Used Terms and Hashtag Analysis

Focusing on English content, the rate of frequently used terms was established, excluding "stop words" (Sinclair and Rockwell, 2016). The terms found were particularly associated with the struggle for human rights, empathy towards the voiceless, and sorrow for the victims of violence. This analysis revealed the primary themes and sentiments expressed by users during the conflict, highlighting the emotional and political dimensions of the discourse.

Hashtag activism evolved on social media to share narratives with like-minded users, creating a sense of community and solidarity around specific issues. 23,109 hashtags were extracted from the corpus of tweets related to the Israeli-Palestinian crisis, with thousands of hashtag variations observed. This diversity of hashtags suggests a complex and multifaceted discourse, with users employing a wide range of tags to express their views and engage with different aspects of the conflict.

Although the #SaveSheikhJarrah movement emerged with high frequency in the first days, its use decreased dramatically in the following days (Figure 2), suggesting that the Sheikh Jarrah evictions were a powerful stimulus for

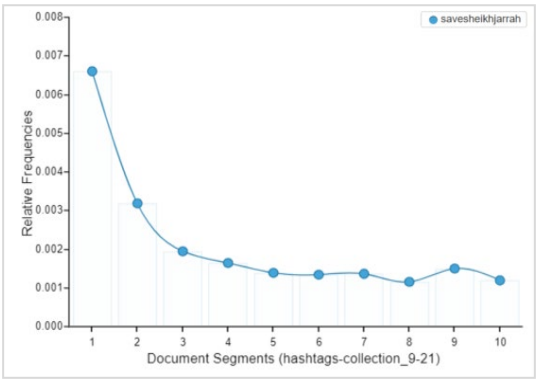


Figure 2: #SaveSheikhJarrah Distribution within the Time

further cybercampaigns rather than a central uniting theme. The decline in the usage of this hashtag over time may indicate a shift in the focus of the discourse or the emergence of new issues and events that captured users' attention.

The analysis demonstrated that adding hashtags to tweets does not necessarily affect their diffusion (figure not included due to space restrictions). Only 13% of the total 370,760 tweets had more than 100 retweets. Within this group, 41% of the most popular tweets contained hashtags, while the remaining 59% did not. The remaining 87% of tweets received less than 100 retweets, indicating that tweets with or without hashtags were retweeted to a similar degree. This finding suggests that the presence of hashtags alone does not guarantee the virality or popularity of a tweet, and other factors, such as content, timing, and the influence of the user, may play a more significant role in determining the reach and impact of a message.

Moving beyond the specific Sheikh Jarrah incident, the remaining hashtags were systematized and presented in a word cloud showing the most used hashtag variations (not included due to space restrictions). Voyant Tools was used to create a word cloud demonstrating time-based changes in hashtag usage, providing a visual representation of the shifting focus and intensity of the discourse over time.

The trends graph revealed a peak in the usage of #GazaUnderAttack in the third corpus segment, corresponding to the peak user activity on May 12. This peak suggests a significant increase in user engagement and concern regarding the situation in Gaza, likely in response to specific events or developments on that day. The graph provided valuable insights into the most frequently used hashtags within the corpus, allowing for a more

nuanced understanding of the key themes and issues driving the conversation.

Although hashtags did not directly affect virality, they provided valid data about the discourse and served as indicators of public opinion and sentiment. The most frequent hashtags extracted were more reliable than infrequent ones, as they represented the views and experiences of a larger number of users. Figures 3 and 4 compare semantically equivalent hashtag groups, revealing patterns of attribution and blame in the context of the conflict. Figure 3 shows the public opinion distribution regarding airstrike responsibility across corpus segments. #IsraeliTerrorism peaked on the night of May 11, when protests were widespread and the first airstrike on Hanadi Tower in Gaza occurred ("Timeline of the Israeli–Palestinian conflict in 2021," 2022). This peak indicates a strong public reaction and condemnation of Israeli actions, with users employing the hashtag to express their outrage and attribute responsibility for the escalation of violence.

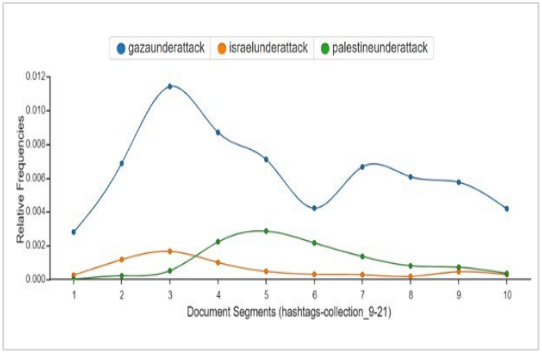


Figure 3: #GazaUnderAttack, #IsraelUnderAttack, #PalestineUnderAttack Distribution Comparison

Exploring Figure 4, specifically segment 3, reveals a drastic increase in #GazaUnderAttack usage, while #IsraelUnderAttack also increased, albeit with a lower frequency. This disparity suggests a stronger focus on the impact of the conflict on Gaza and its residents, with users expressing solidarity and concern for the Palestinian population. The second peak for #GazaUnderAttack appeared on the night between May 17 and 18 when the crossfires resumed (Al-Mughrabi et al., 2021), indicating a renewed surge in user engagement and reaction to the escalating violence. Interestingly, #IsraelUnderAttack usage did not increase in segment 7, despite the ongoing conflict. This may suggest a shift in the discourse or a lack of significant events or developments perceived as threats to Israel during that period. The growth in the #PalestineUnderAttack hashtag observed in the fifth segment coincides with the IDF's targeting of the al-Jalaa building in Gaza ("Timeline of the Israeli–Palestinian conflict in 2021", 2022), indicating a strong public response and condemnation of the attack.

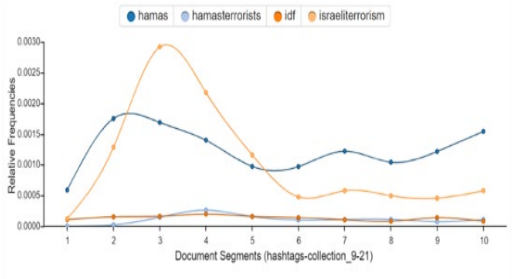


Figure 4: #Hamas, #HamasTerrorist, #IDF, #IsraelTerorist Distribution Comparison

Despite Hamas responding with rockets (Figure 5), there is no evidence of a corresponding increase in the usage of #HamasTerrorists. This finding suggests that public opinion and sentiment, as expressed through hashtags, did not necessarily mirror the military actions of Hamas, and users may have focused on other aspects of the conflict or refrained from attributing terrorist labels to the group. Figure 5 comparatively illustrates two similar but opposing hashtags expressing support for either Israel (#IStandWithIsrael) or Palestine (#IStandWithPalestine). Both trend lines are remarkably fluctuating, with Israeli support rising at the beginning of militant operations (segment 2), possibly indicating a rally-around-the-flag effect or a surge in pro-Israel sentiment in response to the initial escalation of violence. However, Palestinian supporters became prominent from segment 4 onwards, suggesting a shift in public opinion and a growing wave of solidarity with the Palestinian cause. This trend continued until pro-Israeli users re-established their quantitative superiority in the last segment, which may reflect a change in the discourse or a reaction to specific events or developments in the final days of the conflict.

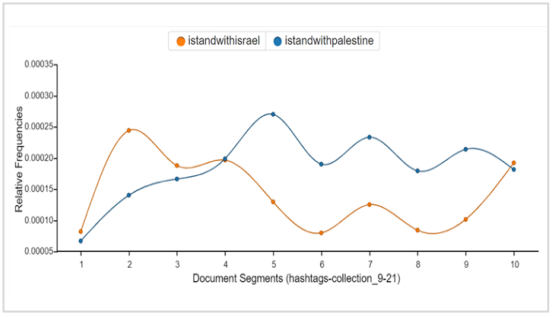


Figure 5: #IStandWithIsrael and #IStandWithPalestine Distribution Comparison

In addition to utilitarian hashtags used by the media to highlight news (#breaking, #Israel,

#Hamas, #Palestine), emotional hashtags played a significant role in the discourse. Hashtags such as #PalestineBleeding, #hearGaza, and #PalestineWillBeFree indicated users' emphatic interest in expressing their feelings and experiences related to the events, often conveying a sense of solidarity, compassion, and support for the Palestinian people.

Despite a sharp decrease in the use of the #SaveSheikhJarrah hashtag after the first airstrikes, it remained among the top hashtags throughout the analyzed period. This sustained presence suggests that while the focus of the discourse may have shifted to other aspects of the conflict, the Sheikh Jarrah evictions continued to be a significant underlying issue and a symbol of the broader struggle for Palestinian rights and self-determination.

Overall, the analysis of frequently used terms and hashtags provides a rich and nuanced understanding of the key themes, sentiments, and dynamics of the Twitter discourse surrounding the May 2021 Israeli-Palestinian conflict. By combining quantitative and qualitative insights, this study sheds light on the complex interplay between public opinion, media coverage, and the evolving narratives and frames employed by users to make sense of the unfolding events.

5. Limitations and Future Directions

Our study's focus on English-language tweets may overlook crucial Arabic and Hebrew perspectives directly involved in the Israeli occupation of Palestine. To address this limitation, future work should adopt a multilingual approach, integrating Arabic and Hebrew tweets and leveraging advanced NLP tools for translation, sentiment analysis, and dialect identification. This will provide a more representative analysis of the diverse narratives surrounding the conflict.

5.1 Methodological Choices

Our methodological framework, particularly keyword selection and data collection timing, may have influenced the research outcomes. The initial keyword selection, guided by prevalent hashtags, may have limited our dataset to predominant narratives, overlooking emerging or nuanced voices. Similarly, the temporal boundary of May 9-21, 2021, could have overlooked insights from the conflict's prelude or subsequent developments. Future work will examine these choices more rigorously, exploring alternative data collection periods, keyword strategies, and supplementary data sources to enhance the analysis's comprehensiveness and representativeness.

5.2 Addressing Interpretative Biases

The polarized nature of the Israeli-Palestinian discourse and the conflict's complexities necessitate examining potential interpretative biases in our analysis. To bolster objectivity and reliability, we will implement strategies such as triangulation, peer debriefing, and member checking. We will also engage in a more explicit discussion of researchers' positionality and reflexivity, acknowledging how our backgrounds and perspectives may have influenced the analysis. By addressing these aspects transparently, we aim to enhance the credibility and trustworthiness of our research insights.

6. Conclusion

This study investigated the use of social media during the May 2021 Israeli-Palestinian crisis, evaluating the hypothesis that social media platforms can effectively trace crisis dynamics. The research fills a gap in the literature by integrating empirical findings with existing theoretical frameworks. The analysis revealed that hashtags like #SaveSheikhJarrah indicated significant shifts in the crisis narrative, and the study extended its focus beyond sentiment analysis to include contextual dynamics.

The findings confirmed social media's utility in crisis tracing, although content virality did not directly correlate with informational substance.

Celebrities and influencers played a significant role in shaping public opinion, while Israeli state-affiliated accounts exerted influence by advocating for IDF operations. Hashtag analytics revealed prevailing pro-Palestinian sentiment and provided insights into public opinion and crisis stages.

Despite limitations in data collection timing, keyword selection, and linguistic constraints, the study documented the evolution of the May 2021 crisis and identified a significant cybercampaign centered around #SaveSheikhJarrah, which engaged a global audience and prominent influencers, revealing polarized support for the conflicting parties.

Future research should focus on the granular analysis of location-based tweet patterns and individual hashtags to enhance the understanding of social media's role in crisis communication.

7. Acknowledgements

This publication was made possible by NPRP14C-0916-210015 / MARSAD Sub-Project from the Qatar National Research Fund / Qatar Research Development and Innovation Council (QRDI). The contents herein reflect the work and are solely the authors' responsibility.

8. References

Alam, F., Barrón-Cedeño, A., Cheema, G. S., Hakimov, S., Hasanain, M., Li, C., Míguez, R., Mubarak, H., Shahi, G. K., Zaghouani, W., & others. (2023). Overview of the CLEF-2023 CheckThat! lab task 1 on check-worthiness in multimodal and multigenre content. Working Notes of CLEF.

Alam, F., Shaar, S., Dalvi, F., Sajjad, H., Nikolov, A., Mubarak, H., Martino, G. D. S., Abdelali, A., Durrani, N., Darwish, K., & others. (2021). Fighting the COVID-19 infodemic: modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. Findings of EMNLP-2021.

Al-Mughrabi, N., Farrell, S., & Heller, J. (2021, May 18). World powers urge truce as Israel-Palestinian conflict rages. Reuters. https://cutt.ly/lObLy8b

An, J., Quercia, D., Cha, M., Gummadi, K., & Crowcroft, J. (2014). Sharing political news: The balancing act of intimacy and socialization in selective exposure. EPJ Data Science, 3, 1-21. https://doi:10.1140/epjds/s13688-014-0012-2

Anthony, L. (2010). AntConc (Version 4.0.3.0) [Computer Software]. http://www.antlab.sci.waseda.ac.jp/

Alimardani, M., & Elswah, M. (2021, August 5). Digital orientalism: #SaveSheikhJarrah and Arabic content moderation. POMEPS Studies, 43, 69-75. https://ssrn.com/abstract=3900520

Atanasova, P., Barron-Cedeno, A., Elsayed, T., Suwaileh, R., Zaghouani, W., Kyuchukov, S., Martino, G. D. S., & Nakov, P. (2018). Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims. Task 1: Check-worthiness. ArXiv Preprint ArXiv:1808.05542.

Azzimonti, M., & Fernandes, M. (2018). Social media networks, fake news, and polarization. National Bureau of Economic Research. https://www.nber.org/papers/w24462

Bodaghi, A., & Oliveira, J. (2020). The characteristics of rumor spreaders on Twitter: A quantitative analysis on real data. Computer Communications, 160, 674-687. https://doi.org/10.1016/j.comcom.2020.07.017

Ajao, O., Bhowmik, D., & Zargari, S. (2018, July). Fake news identification on Twitter with hybrid CNN and RNN models. Proceedings of the 9th International Conference on Social Media and Society, 226-230. https://doi.org/10.1145/3217804.3217917

Al Sarraj, W. F., Kahloot, K. M., Maghari, A. Y., & Abu-Ghosh, M. M. (2016, August). A social network analysis of tweets during the Gaza War, summer 2014. 2016 IEEE 4th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW), 220-227. https://doi.org/10.1109/W-FiCloud.2016.54

Barkho, L., & Richardson, J. (2010). A critique of BBC's Middle East news production strategy. American Communication Journal, 12(1). http://urn.kb.se/resolve?urn=urn:nbn:se:hj:diva-17394

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative research in Psychology, 3(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Cacciatore, M. A., Scheufele, D. A., & Iyengar, S. (2016). The end of framing as we know it… and the future of media effects. Mass Communication and Society, 19(1), 7-23. https://doi.org/10.1080/15205436.2015.1068811

Ciszek, E. L. (2016). Digital activism: How social media and dissensus inform theory and practice. Public Relations Review, 42(2), 314-321. https://doi.org/10.1016/j.pubrev.2016.02.002 .

Deetz, S. (2001). Conceptual foundations. In F. M. Jablin & L. L. Putnam (Eds.), The New Handbook of Organizational Communication (pp. 3-46). SAGE Publications, Inc., https://dx.doi.org/10.4135/9781412986243

Du, S., & Gregory, S. (2016, November). The echo chamber effect in Twitter: Does community polarization increase? International Workshop on Complex Networks and Their Applications, 373-378. https://doi.org/10.1007/978-3-319-50901-3_30

Fekete, E., & Warf, B. (2013). Information technology and the “Arab Spring”. The Arab World Geographer, 16(2), 210-227. https://doi.org/10.5555/arwg.16.2.u2q0427u4883l635

Felt, M. (2016). Social media and the social sciences: How researchers employ Big Data analytics. Big Data & Society, 3(1). https://doi.org/10.1177%2F2053951716645828

Garimella, K., De Francisci Morales, G., Gionis, A., & Mathioudakis, M. (2018, April). Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. Proceedings of the 2018 World Wide Web Conference, 913-922. https://doi.org/10.1145/3178876.3186139

Golkar, S. (2011). Liberation or suppression technologies? The Internet, the Green Movement and the regime in Iran. International Journal of Emerging Technologies & Society, 9(1), 50-70. https://cutt.ly/UU6KjyG

Grömping, M. (2014). ‘Echo chambers’ partisan Facebook groups during the 2014 Thai election. Asia Pacific Media Educator, 24(1), 39-59. https://doi.org/10.1177%2F1326365X14539185

Hasanain, M., Alam, F., Mubarak, H., Abdaljalil, S., Zaghouani, W., Nakov, P., Martino, G. D. S., & Freihat, A. A. (2023). Araieval shared task: Persuasion techniques and disinformation detection in arabic text. ArXiv Preprint ArXiv:2311.03179

Holtzhausen, D. R. (2002). Towards a postmodern research agenda for public relations. Public Relations Review, 28(3), 251-264. https://doi.org/10.1016/S0363-8111(02)00131-5

Holtzhausen, D. R., & Voto, R. (2002). Resistance from the margins: The postmodern public relations practitioner as organizational activist. Journal of Public Relations Research, 14(1), 57-84. https://doi.org/10.1207/S1532754XJPRR1401_3

Hoskins, A., & O'Loughlin, B. (2015). Arrested war: The third of mediatization. Information, Communication & Society, 18(11), 1320-1338. http://dx.doi.org/10.1080/1369118X.2015.1068350

Huang, Y. L., Starbird, K., Orand, M., Stanek, S. A., & Pedersen, H. T. (2015, February). Connected through crisis: Emotional proximity and the spread of misinformation online. Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, 969-980. https://doi.org/10.1145/2675133.2675202

Hunt, K., Wang, B., & Zhuang, J. (2020). Misinformation debunking and cross-platform information sharing through Twitter during Hurricanes Harvey and Irma: A case study on shelters and ID checks. Natural Hazards, 103, 861-883. https://doi.org/10.1007/s11069-020-04016-6

Jain, S., Sharma, V., & Kaushal, R. (2016, September). Towards automated real-time detection of misinformation on Twitter. 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2015-2020. https://doi.org/10.1109/ICACCI.2016.7732347

Jamison, A., Broniatowski, D. A., Smith, M. C., Parikh, K. S., Malik, A., Dredze, M., & Quinn, S. C. (2020). Adapting and extending a typology to identify vaccine misinformation on Twitter. American Journal of Public Health, 110(S3), 331-339. https://doi.org/10.2105/AJPH.2020.305940

Kent, M. L. (2013). Using social media dialogically: Public relations role in reviving democracy. Public Relations Review, 39(4), 337-345. https://doi.org/10.1016/j.pubrev.2013.07.024

Kušen, E., & Strembeck, M. (2018). Politics, sentiments, and misinformation: An analysis of the Twitter discussion on the 2016 Austrian Presidential Elections. Online Social Networks and Media, 5, 37-50. https://doi.org/10.1016/j.osnem.2017.12.002

Laabar, S., & Zaghouani, W. (2024). Multi-dimensional insights: Annotated dataset of stance, sentiment, and emotion in Facebook comments on Tunisia's July 25 measures. In Proceedings of the Second Workshop on Natural Language Processing for Political Sciences* co-located with the 2024 International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024).

Lawrence, E., Sides, J., & Farrell, H. (2010). Self-segregation or deliberation? Blog readership, participation, and polarization in American politics. Perspectives on Politics, 8(1), 141-157. https://doi:10.1017/S1537592709992714

Lewandowsky, S., Oberauer, K., & Gignac, G. E. (2013). NASA faked the moon landing – therefore, (climate) science is a hoax: An anatomy of the motivated rejection of science. Psychological Science, 24(5), 622-633. https://doi.org/10.1177%2F0956797612457686

Manor, I., & Crilley, R. (2018). Visually framing the Gaza War of 2014: The Israel ministry of foreign affairs on Twitter. Media, War & Conflict, 11(4), 369-391. https://doi.org/10.1177/1750635218780564

Martínez García, A. B. (2017). Bana Alabed: Using Twitter to draw attention to human rights violations. Prose Studies, 39(2-3), 132-149. https://doi.org/10.1080/01440357.2018.1549310

McGlynn, J., Baryshevtsev, M., & Dayton, Z. A. (2020). Misinformation more likely to use non-specific authority references: Twitter analysis of two COVID-19 myths. Harvard Kennedy School Misinformation Review, 1(3). https://doi.org/10.7910/DVN/GSFFFP

Mtchedlidze, J. (2019). A discourse analysis of war representation on Twitter by civilian actors. A case of the Gaza-Israel war in 2014

Mueller, A., Wood-Doughty, Z., Amir, S., Dredze, M., & Nobles, A. L. (2021). Demographic representation and collective storytelling in the Me Too Twitter hashtag activism movement. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), 1-28. https://doi.org/10.1145/3449181

Naeem, M., & Ozuem, W. (2021). Understanding misinformation and rumors that generated panic buying as a social practice during COVID-19 pandemic: Evidence from Twitter, YouTube and focus group interviews. Information Technology & People. https://doi.org/10.1108/ITP-01-2021-0061

Rangel, F., Rosso, P., Charfi, A., Zaghouani, W., Ghanem, B., & Sánchez-Junquera, J. (2019). Overview of the track on author profiling and deception detection in arabic. Working Notes of FIRE 2019. CEUR-WS. Org, Vol. 2517, 70–83.

Rosso, P., Rangel, F., Farías, I. H., Cagnina, L., Zaghouani, W., & Charfi, A. (2018). A survey on author profiling, deception, and irony detection for the arabic language. Language and Linguistics Compass, 12(4), e12275.

Sanderson, Z., Brown, M. A., Bonneau, R., Nagler, J., & Tucker, J. A. (2021). Twitter flagged Donald Trump's tweets with election misinformation: They continued to spread both on and off the platform. Harvard Kennedy School Misinformation Review. https://doi.org/10.7910/DVN/DDJNEF

Seo, H. (2014). Visual propaganda in the age of social media: An empirical analysis of Twitter images during the 2012 Israeli-Hamas conflict. Visual Communication Quarterly, 21(3), 150-161. https://doi.org/10.1080/15551393.2014.955501

Shaar, S., Alam, F., Martino, G. D. S., Nikolov, A., Zaghouani, W., Nakov, P., & Feldman, A. (2021). Findings of the NLP4IF-2021 shared tasks on fighting the COVID-19 infodemic and censorship detection. ArXiv Preprint ArXiv:2109.12986.

Shahi, G. K., Dirkson, A., & Majchrzak, T. A. (2021). An exploratory study of COVID-19 misinformation on Twitter. Online Social Networks and Media, 22. https://doi.org/10.1016/j.osnem.2020.100104

Shaw, A. (2012). Centralized and decentralized gatekeeping in an open online collective. Politics & Society, 40(3), 349-388. https://doi.org/10.1177%2F0032329212449009

Shehabat, A. (2012). The social media cyber-war: The unfolding events in the Syrian revolution 2011. Global Media Journal: Australian Edition, 6(2). http://handle.uws.edu.au:8081/1959.7/538867

Shreim, N., & Dawes, S. (2015). Mediatizing Gaza: An introduction. Networking Knowledge: Journal of the MeCCSA Postgraduate Network, 8(2). https://doi.org/10.31165/nk.2015.82.367

Shurafa, C., Darwish, K., & Zaghouani, W. (2020). Political framing: US COVID19 blame game. Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, Proceedings 12, 333–351.

Simonsen, S. (2019). Discursive legitimation strategies: The evolving legitimation of war in Israeli public diplomacy. Discourse & Society, 30(5), 503-520. https://doi.org/10.1177/0957926519855786

Sinclair, S., & Rockwell, G. (2016). Voyant Tools. http://voyant-tools.org/

Stein, R. L. (2012). StateTube: Anthropological reflections on social media and the Israeli state. Anthropological Quarterly, 85(3), 893-916. http://www.jstor.org/stable/41857275

Tawil-Souri, H., & Aouragh, M. (2014). Intifada 3.0? Cyber colonialism and Palestinian resistance. The Arab Studies Journal, 22(1), 102-133. https://www.jstor.org/stable/24877901

Taylor, M., & Kent, M. L. (2010). Anticipatory socialization in the use of social media in public relations: A content analysis of PRSA's public relations tactics. Public Relations Review, 36(3), 207-214. https://doi.org/10.1016/j.pubrev.2010.04.012

Wallace, J. (2018). Modelling contemporary gatekeeping: The rise of individuals, algorithms and platforms in digital news dissemination. Digital Journalism, 6(3), 274-293. https://doi.org/10.1080/21670811.2017.1343648

Zaghouani, W., & Charfi, A. (2018). Arap-tweet: A large multi-dialect twitter corpus for gender, age and language variety identification. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Zeitzoff, T. (2011). Using social media to measure conflict dynamics: An application to the 2008–2009 Gaza conflict. Journal of Conflict Resolution, 55(6), 938-969. https://doi.org/10.1177%2F0022002711408014

Zeitzoff, T. (2017). How social media is changingconflict. Journal of Conflict Resolution, 61(9), 1970-1991. https://doi.org/10.1177%2F0022002717721 39

Zeitzoff, T. (2018). Does social media influence conflict? Evidence from the 2012 Gaza Conflict. Journal of Conflict Resolution, 62(1), 29-63. https://doi.org/10.1177%2F0022002716650925

Zeitzoff, T., Kelly, J., & Lotan, G. (2015). Using social media to measure foreign policy dynamics: An empirical analysis of the Iranian-Israeli confrontation (2012–13). Journal of Peace Research, 52(3), 368-383. https://doi.org/10.1177%2F0022343314558700

Semi-Automatic Topic Discovery and Classification for Epidemic Intelligence via Large Language Models

Federico Borazio[†], Danilo Croce[†], Giorgio Gambosi[†], Roberto Basili[†], Daniele Margiotta[‡], Antonio Scaiella[‡], Martina Del Manso[*], Daniele Petrone[*], Andrea Cannone[*], Alberto Mateo Urdiales[*], Chiara Sacco[*], Patrizio Pezzotti[*], Flavia Riccardo[*], Daniele Mipatrini[+], Federica Ferraro[+], Sobha Pilati[+]

[†] Department of Enterprise Engineering, University of Rome Tor Vergata, Italy

[‡] Reveal s.r.l.

[*] Infectious Diseases Department - Istituto Superiore della Sanità

[+] General Directorate for Health Prevention - Italian Ministry of Health

borazio@ing.uniroma2.it, {croce,basili}@info.uniroma2.it

Abstract

This paper introduces a novel framework to harness Large Language Models (LLMs) for Epidemic Intelligence, focusing on identifying and categorizing emergent socio-political phenomena within health crises, with a spotlight on the COVID-19 pandemic. Our approach diverges from traditional methods, such as Topic Models, by providing explicit support to analysts through the identification of distinct thematic areas and the generation of clear, actionable statements for each topic. This supports a Zero-shot Classification mechanism, enabling effective matching of news articles to fine-grain topics without the need for model fine-tuning. The framework is designed to be as transparent as possible, producing linguistically informed insights to make the analysis more accessible to analysts who may not be familiar with every subject matter of inherently emerging phenomena. This process not only enhances the precision and relevance of the extracted Epidemic Intelligence but also fosters a collaborative environment where system linguistic abilities and the analyst's domain expertise are integrated.

Keywords: Epidemic Intelligence, Topic Discovery, Large Language Models, Zero-shot Classification

1. Epidemic Intelligence: Objectives and Challenges

Following the paradigmatic change from disease specific to an all-hazard approach to the assessment of public health introduced in the 2005 revision of the International Health Regulations[1], the concept of Epidemic Intelligence was defined as a complex of activities related to the early identification of potential health hazards, their verification, assessment, and investigation that aim to generate information to guide appropriate actions in public health (Paquet et al., 2006), (World Health Organization, 2014). Within this global framework, Member States have developed ways to implement this concept to support situation awareness and evidence-based decision-making in public health. Italy started to develop its own national approach to Epidemic Intelligence in 2007 as part of a project funded by the Italian Ministry of Health coordinated by the Istituto Superiore di Sanità (ISS) (Del Manso et al., 2022). At this time a situation and need assessment was performed in order to identify existing capacities and areas with additional implementation requirements.

The results led to the conclusion that while the epidemiological monitoring conducted on data generated by existing national surveillance systems for infectious diseases (clinical, laboratory-based, and syndromic) could support an indicator-based component for the early detection of transmission events in the country, an Epidemic Intelligence system in Italy would need to develop ex novo an event based surveillance component. This component would be an extremely sensitive and flexible surveillance system based on open-source unstructured information published online concerning cases and clusters of infectious disease occurring in Italy in order to inform as soon as possible decision-making and public health experts or to provide information to clinicians and improve the timeliness of diagnoses. Some of this information would be validated (i.e. verified with public health officials within the country). The selection and assessment of news items would be performed by trained analysts to detect events of public health importance according to the methodology developed by the European Centre for Disease Prevention and Control (ECDC[2]). Following several pilots to design and test this national event-based surveillance component of Epidemic Intelligence, Italy chose to follow the implementation model developed and sustain-

[1] International Health Regulations (2005): https://iris.who.int/bitstream/handle/10665/43883/9789241580410_eng.pdf

[2] ECDC: https://www.ecdc.europa.eu/en/news-events/e-learning-course-epidemic-intelligence-ei

Proceedings of the Second Workshop on Natural Language Processing for Political Sciences (PoliticalNLP 2024), pages 68–84

ably implemented by the Global Health Security Action Group Early Alerting and Reporting project (EAR) (Riccardo et al., 2014). This consisted of a decentralized approach in which participating countries contributed analysts that were operational on a rotation basis.

In order to apply this to the Italian regionalized health care system, since 2017, Italy has adopted a decentralized method of setting up a network of analysts (Network Italiano di Epidemic Intelligence - Italian Network of Epidemic Intelligence) nominated by regional authorities among subject-matter experts employed within the national health system at the national, regional and local level. Analysts of the Italian Network of Epidemic Intelligence work in rotating teams. Each day they screen news items, identifying those that are relevant to the surveillance focus (e.g., cases or clusters of infectious diseases in Italy or/and signs and symptoms in unexpected frequency) that are called signals. Signals are then individually risk assessed by the analysts using a common methodology (Intelligence and Miglietta, 2022) to identify those of public health relevance that are called events and that are then reported. At any given time, analysis is required to manually screen thousands of news items, reject irrelevant ones categorize signals, and assess them as events. Especially the screening phase of this work is extremely time-consuming and resource intensive and this undermines the long-term sustainability of this surveillance system. The integration of NLP techniques brings a significant contribution to the quality enhancement of the Epidemic Intelligence efforts: (*i*) *Enhanced Text search capabilities*, enabling the processing of larger text data volumes to uncover emerging threats, thus aiding to identify otherwise overlooked information; (*ii*) *Reduced monitoring time*, through the automation of routine monitoring tasks, allowing to allocate more time to complex and strategic analyses; and (*iii*) *Improved accuracy*, fostering for well-informed and documented decisions.

This paper introduces an advanced framework designed to harness the capabilities of Large Language Models (LLMs) for Epidemic Intelligence, addressing the specific challenges of identifying and categorizing emergent socio-political phenomena within the context of health crises, notably the COVID-19 pandemic. The objective is not to replace the analyst with an opaque, black-box approach for topic discovery but to ensure each analytical step is as self-explanatory as possible. By producing linguistically informed insights, we aim to elucidate the rationale behind the interpretation, for the analysts not familiar with subject matters about inherently emerging phenomena. The methodology begins with the user need to be defined as a set of seed terms to delineate a high-level conceptual document perimeter. This produces a corpus of retrieved news, specifically focused on the general need. Then, linguistic triples are generated to capture fine-grain concepts (e.g. specific activities or clinical concepts) implicit in the corpus. These triples can then be validated manually against the collected news, culminating in their automatic translation into prompts. Notice that the user can continuously fine-tune the system's proposed prompts, further customizing the analysis to his own specific needs. The overall process fosters a collaborative environment where the agent's intelligence and the analyst's domain expertise cooperate according to the investigative goals. This enhances the accuracy and coverage of the resulting Epidemic Intelligence activities. Preliminary results from our empirical investigation confirm the significant benefits of our workflow's capability in accurately mapping news articles to pertinent fine-grain topics. In the remaining, Section 2 reports related work, Section 3 presents the proposed workflow, Section 4 discusses the empirical evaluation, while Section 5 presents the conclusions.

2. Related Work

The use of NLP and text mining techniques in order to extract relevant information from vast amounts of text data available on the internet, thus allowing the identification of relevant epidemiological events, has been extensively studied in the previous years (see (O'Shea, 2017) for a systematic review of proposals dating a few years ago). For instance, the Medical Information System (MedISys) (Rortais et al., 2010) supports the timely detection of emerging diseases by crawling online news articles and applying hierarchical clustering algorithms to classify them into predefined categories. The Pattern-based Understanding and Learning System (PULS) (Yangarber and Steinberger, 2009) extends the MedISys by applying Natural Language Processing. The Global Health Monitor system (Collier et al., 2008) uses instead an ontology-based approach to text mining text data from the web to detect and track infectious disease outbreaks.

Text classification is a fundamental approach to the identification of relevant events. After early works applying classical machine learning approaches (Kowsari et al., 2019) (Khan et al., 2010), deep learning architectures introduced a new set of general methods (Minaee et al., 2021), (Luan and Lin, 2019) for text and news classification. Interest in the topic received a boost with the outbreak of the COVID-19 pandemic (Al-Garadi et al., 2022), (Raza et al., 2022), (Raza and Schwartz, 2023). Moreover, the advent of the attention mechanism in neural networks (Vaswani et al., 2017) and the adoption of transformer-based encoders

(Devlin et al., 2019), (Gillioz et al., 2020) made it possible effective information extraction from texts (Gupta et al., 2022), (Choudhary et al., 2023) as well as document classification (Li et al., 2022), (Deping et al., 2021), (Kaliyar et al., 2021). The use of transformer architectures, and the related Large Language Models, to news classification is an active research area (see for example (Khosa et al., 2023), (Deping et al., 2021), (Santana et al., 2022), (Gunes and Florczak, 2023)). However, the evaluation of the use of such approaches to the medical, and in particular in the epidemiological, field has been performed only quite recently (Wang et al., 2023), (Adaszewski et al., 2021) and it is still in its infancy. Unlike traditional approaches that might rely on probabilistic distributions akin to Topic Models (Blei et al., 2003), (Blei and Lafferty, 2009), (Mcauliffe and Blei, 2007),(Churchill and Singh, 2022), our method aims to provide explicit support to analysts. It does so by identifying distinct thematic areas and generating clear, actionable statements for each topic, such as "*A news article pertains to this topic if it addresses* ...". These statements are straightforward triggers for a Zero-shot Classification mechanism (Yin et al., 2019), effectively matching news articles to meticulously defined topics.

3. Automatic Topic Discovery and Classification

This section describes the workflow from initial data gathering to the application of Zero-shot classifiers for identifying and categorizing emergent socio-political phenomena.

The **Data Gathering** phase initiates our workflow, where analysts input keywords, such as "*Coronavirus outbreak*" or "*Covid contagion*", to guide the system in collecting news articles relevant to the defined scope. This stage aims to cast a wide net to ensure comprehensive coverage, setting the stage for subsequent refinement and analysis.

The cornerstone of our methodology lies in harnessing the initial intuition of analysts to seed the discovery of diverse topics within the vast landscape of collected data. This phase, **Topic Discovery**, begins with the identification of *seed-words*, terms that encapsulate the essence of the analyst's goal. In the Topic Discovery phase, seed words serve as descriptors of a general domain of analysis and let the system suggest potential themes. The analysts can then refine or expand upon these suggestions. This interaction balances automation and human expertise, ensuring a form of both guided and nuanced analysis. Terms such as "*Covid*" and "*Hospital*" could serve as initial seeds, delimiting the text perimeter within the broader domain of public health and epidemic preparedness. The output of this phase is a set of specific concepts that should emerge directly from the news in the perimeter, different from abstract word distribution, often associated with traditional Topic Modeling approaches, (Blei et al., 2003; Abdelrazek et al., 2023). Our target is not generating probabilistic topic models but conceptual sub-topics that are immediately comprehensible to an analyst. In this work, we thus conceive a topic as a collection of assertions such as "*This text discusses a concerning increase in infections led to a rise in Covid patients*." or "*The text discusses the monitoring of the increase in deaths caused by Covid*" for a topic possibly named "Pandemic Peaks". Another such topic as "Covid patient care" could be inspired by assertions like "*The text addresses the challenges hospitals face in accommodating new patients*" or "*The text discusses hospitals continuing to vaccinate patients for COVID prevention*". This approach makes topic interpretation easy, given the assertions and the title. Whether a news article aligns with a specific topic depends upon the verification of its assertions. A news article is assigned to a topic proportionally to how much the system judges one of its assertions to be true. Multiple true assertions incrementally contribute to the overall confidence in the topic. Crucially, as assertions are interpreted first by the analysts, he may wish to refine a topic by adjusting, deleting, or introducing new assertions.

In the final phase, we apply **Zero-shot Classification** through Large Language Models (LLMs) to conclusively label news articles with the specific topics identified earlier, avoiding model fine-tuning. This approach, grounded in natural language inference (Yin et al., 2019), exploits logical alignment between text and topic-defining assertions in the form of prompts. Notice that this enhances article-topic associations, beyond mere categorization, as assertions also provide explanations of individual classification inferences.

Upon completion of the workflow, once topics and corresponding prompts are made available, all news can be classified accordingly. Users can then exploit specific topics, e.g.,"Pandemic Peaks", as news filters, based on the metadata associated with prompts. This enables further analyses, such as focused news retrieval, filtering, and aggregation.

3.1. Data Gathering

The foundation of our approach begins with the **Data Gathering** phase, a crucial step designed to amass a comprehensive corpus of news articles pertinent to specific events or phenomena. For instance, an analyst may conduct an inquiry into the societal impact of afflictions, such as the Coronavirus within the Italian territory over the past fortnight. Accordingly, by providing pivotal terms such as "*Coronavirus outbreak*" and/or "*Covid contagion*"

(possibly accompanied by time constraints) the process autonomously assembles a specific document collection, through the systematic extraction of Web news articles. In the initial phase, broad or generic query terms are used to maximize coverage which means extending article retrieval also to possibly irrelevant data. This strategy aims to extend the corpus, leaving its refining and validating to subsequent, more informed, stages. To facilitate this process, we developed a dedicated crawling service, aimed at collecting unstructured data using Google News as a primary but not exclusive source.

3.2. Topic Discovery

The Topic Discovery phase is pivotal in our workflow, aimed at moving from a small set of seed words to a possibly comprehensive collection of specific epidemic topics. Consider the previous example of an analyst inputting seeds such as "*Covid*" and "*Hospitals*". The initial step of lexical expansion endeavors to broaden the analyst's query using Word Space models, such as those created by the Contextual Bag of Words (CBOW) model implemented in `Word2Vec` (Mikolov et al., 2013). This distributional representation embeds terms within high-dimensional spaces where metric distances mirror paradigmatic relations, like quasi-synonymy, facilitating the exploration of related lexical fields (Sahlgren, 2006). Expanding upon the initial seed terms involves selecting the terms closest to each seed, aiming to broaden the initial semantics for topic generation. For example, the most similar words to "*Hospital*" are "*clinic*" and "*infirmary*", while "*coronavirus*" and "*pandemic*" are the corresponding words for "*Covid*". These entries offer a useful semantic expansion for the analysts to explore related themes. However, complex prompts for classification (i.e. assertions) require more informative linguistic structures corresponding to concepts, such as biological or clinical entities or events. This requires not just the selection of individual relevant terms but also complex well-formed definitions.

In this work, to automatically discover meaningful statements, such as "*This text discusses a concerning increase in infections led to a rise in Covid patients*.", we employ a form of grammatically controlled lexical expansion. From seed terms, we aim to generate structured forms like Subject-Verb-Object (SVO) triples, which can be easily transformed into coherent sentences. This approach ensures that an expansion can be easily understood by the analysis, but also facilitates the automatic creation of meaningful textual prompts. We call this step the **Linguistic Triple Generation** process. Assuming the inserted seeds are nouns, our process begins by identifying the set of e_v verbs closest to them. We use cosine similarity in the employed Word Space, also assuming that a seed noun can function either as a subject or as an object of the selected verb. For each such verb v, we then in turn retrieve the set of e_n nouns closest to v to completely fill an SVO structure. This approach ensures that the expansion from seed words to SVO triples is both deliberate and meaningful, providing a semantically rich lexicon from which complex thematic prompts can be derived. For instance, from the seed "*Covid*", we may derive closely related verbs such as "*record*" and "*infect*", while "*hospital*" might lead us to "*admit*" or "*vaccinate*". The expansion from verbs to nouns allows for the generation of SVO triples by further associating these verbs with relevant nouns: "*infect*" leads to "*patients*", "*lung*", and "*infections*"; "*record*" to "*deaths*", "*recovered*", and "*amount*". These expansions facilitate the construction of SVO triples such as ("*Covid*", "*record*", "*deaths*"), ("*Covid*", "*record*", "*recovered*"), ("*Covid*", "*increase*", "*infections*"), ("*Covid*", "*infect*", "*patients*"), ("*Covid*", "*infect*", "*lung*"), ("*Covid*", "*record*", "*infections*"), ("*Hospitals*", "*admit*", "*patients*"), ("*Hospitals*", "*vaccinate*", "*patients*"), ("*Covid*", "*record*", "*amount*"), and ("*Covid*", "*increase*", "*quantity*"), ...

Obviously, the growth of the number of triples given e_s seeds alongside e_v verbs and e_n nouns impacts significantly on complexity. However, the news collection provided by the Data Gathering phase is crucial in filtering triples whose frequency in the corpus is too low, e.g. below a threshold of τ sentences. We can thus manage the proliferation of triples. This approach integrates the semantics of the wordspace with distributional information related to the topics implicitly expressed by the gathered collection.

After the generation of SVO triples, the workflow progresses to **Triple Clustering**, a crucial step designed to detect k distinct thematic areas relevant to the analyst's interests. Notice that each triple is defined in a metric space depending on its Compositional Distributional Semantics (Mitchell and Lapata, 2008). By representing triples as centroids of their constitutive vectors we may map triples in the same wordspace as the lexical entries. This representation supports the clustering of triples whereas the compositional nature of the mapping is useful to preserve semantic proximity, i.e. relatedness between triples. The clustering (via a k-mean-like algorithm) operates in the structured space and induces coherent groups. Each cluster emerges as a thematic entity, characterized by a given unique narrative thread, but described by the semantic proximity among its constituent triples. As an example, applying k-means with $k = 3$ to the SVO triples mentioned above, we derive the following groups: C_1 = {("*Covid*", "*increase*", "*infections*"), ("*Covid*", "*record*", "*deaths*"), ("*Covid*", "*record*", "*recovered*")}, which encapsu-

lates the theme of the increase in Covid-related infections. C_2 = {("*Hospitals*", "*admit*", "*patients*"), ("*Hospitals*", "*vaccinate*", "*patients*"), ("*Covid*", "*infect*", "*patients*"), ("*Covid*", "*infect*", "*lung*"), ("*Hospital*", "*treat*", "*cure*")}, focusing on hospital responses to Covid, including admissions, vaccinations, and treatments. C_3 = {("*Covid*", "*increase*", "*quantity*"), ("*Covid*", "*record*", "*amount*")}, centering on quantitative aspects of the Covid pandemic.

In the provided toy example, the number of triples and clusters is modest, but one can easily envision scenarios with significantly larger outcomes. Moreover, given the limited textual context, triples may emerge redundantly within a given cluster. However triple similarity in the metric space (modeling for example too similar subjects or objects across triples) can be used to automatize a further stage, called **Triple Pruning**. It ensures the satisfaction of some constraints onto triples: (*i*) each triple must be locally informative (provide high levels of *inner novelty*), (*ii*) triples within the same cluster must exhibit large diversity (*outer novelty*), and (*iii*) triples must be *relevant* within the collection of retrieved documents. Ranking triples refines clustering results, enhancing topic specificity and relevance.

Formally, we define a generic triple of terms such that $t_i = (\mathcal{S}_i, \mathcal{V}_i, \mathcal{O}_i)$ where $\mathcal{S}$ denotes the subject, $\mathcal{V}$ denotes the verb and $\mathcal{O}$ denotes the object, each represented by a corresponding embedding vector $\mathbf{s}_i$, $\mathbf{v}_i$, $\mathbf{o}_i$ in a normalized space, i.e., $\|\mathbf{s}_i\| = \|\mathbf{v}_i\| = \|\mathbf{o}_i\| = 1$. After the system has generated the clusters, the selection procedure of best informative triples within a cluster C_j ($j \in 1, \ldots, m$) is set up by combining the semantic signal provided by the documents and the terms of the triples.

First of all, we introduce a cluster such that $C = \{(t_i, w_i)\}$, where w_i is a semantic weighting function for t_i that will be hereafter defined. In order to pick up the triples that provide additional semantic information, we introduce the `Inner Novelty` with the aim of selecting only triples exhibiting meaningful signals through higher internal information heterogeneity. Let

$$in(t_i) = in(\mathcal{S}_i, \mathcal{V}_i, \mathcal{O}_i) = \\ = 1 - \Big(\beta^{\mathcal{SV}}(\mathbf{s_i} \cdot \mathbf{v_i}) + \beta^{\mathcal{SO}}(\mathbf{s}_i \cdot \mathbf{o}_i) + \beta^{\mathcal{VO}}(\mathbf{v}_i \cdot \mathbf{o}_i)\Big)$$

with $\beta^{\mathcal{SV}}, \beta^{\mathcal{SO}}, \beta^{\mathcal{VO}} \in \Re^+$, such that $\beta^{\mathcal{SV}} + \beta^{\mathcal{SO}} + \beta^{\mathcal{VO}} = 1$. In this scenario, we postulate that a triple such as ("*Hospital*", "*treat*", "*cure*") might exhibit low *Inner novelty*, contributing minimally to the analysis due to the high similarity between "*treat*" and "*cure*". The object in this case adds little to the action's significance, leading us to consider its utility in the analysis as marginal. An additional measure is the `Outer Novelty` that captures the relevance of the semantic signal provided by a triple, evaluating the diversity between pairs of triples. Let

$$nov(t_i, t_h) = nov\Big((\mathcal{S}_i, \mathcal{V}_i, \mathcal{O}_i), (\mathcal{S}_h, \mathcal{V}_h, \mathcal{O}_h)\Big) \\ = 1 - \Big(\gamma^{\mathcal{S}}(\mathbf{s}_i \cdot \mathbf{s}_h) + \gamma^{V}(\mathbf{v}_i \cdot \mathbf{v}_h) + \gamma^{O}(\mathbf{o}_i \cdot \mathbf{o}_h)\Big)$$

and with $\gamma^S, \gamma^V, \gamma^O \in \Re^+$ that regulate the *term-wise similarity*, of pairs of triples, with and $\gamma^S + \gamma^V + \gamma^O = 1$. Then, we compute the *Outer Novelty* of a triple relative to a set C of other triples already selected such that:

$$on_i(C) = \begin{cases} \min_{t_h \in C} \ nov(t_i, t_h) & \text{if } C \neq \varnothing \\ 1 & \text{otherwise} \end{cases}$$

where $i \neq h$ and C is the set that contains already chosen triples. For example, against the cluster $C = \{$("*Hospital*", "*treat*", "*case*")$\}$, we hypothesize that a triple such as ("*Hospital*", "*treat*", "*patient*") has a lower *Outer Novelty* when it is less frequent in the collected news corpus than the member of C.

Given a cluster C made of triples ranked according to the defined novelty weights, the overall weight $w_i(C)$ of a triple t_i is:

$$w_i(C) = \log df_i \cdot in_i \cdot on_i(C) \qquad (1)$$

where $\log df_i$ denotes the logarithm of triple's *document frequency*.

The following algorithm in 1 computes the target set C^* of the most informative tuples from a set C, i.e.

$$C^* = \textsc{bestTriples}(C) \subseteq C$$

Obviously, $C^* = \varnothing$ is the initial set of selected triples.

Every generic element $x_i \in C$ corresponds to a 4-tuple

$$x_i = \langle t_i, \log df_i, in_i, on_i(C) \rangle$$

where at the beginning $\forall x_i \in C \ on_i = 1$ as $C^* = \varnothing$, and $w_i(C^*) = \log df_i \cdot in_i$.

Algorithm 1 Selection of best triples

procedure BESTTRIPLES(C)
  $C^* \leftarrow \varnothing$
  $R = \{x_i \in C \mid \log df_i > \tau\}$
  while $R \neq \varnothing$ **do**
    $x \leftarrow \text{argmax}_{\forall x_i \in R}, w_i(C^*)$
    $C^* \leftarrow C^* \cup \{x\}$
      ▷ pruning the less informative triples
        ▷ according to *Outer Novelty*
    $R = \{x_i \in R \mid t_i \neq t \wedge$
      $\min(on_i(C^*), nov(t_i, t)) > \epsilon\}$
  end while
  return C^*
end procedure

Notice that τ and ϵ as the two parameters of the algorithm: τ is the lower bound of frequencies needed to discard too rare triples that are not relevant the a news collection, while $\epsilon \in [0, 1]$ regulate the overall novelty of a new triple against the already selected ones. Moreover, selecting the $\min(on_i(C^*), nov(t_i, t))$ requires constant time thanks to caching. In fact, $\forall x_i \in R$, $on_i(C^*)$ changes at each step as $x_i = \langle t_i, \log df_i, in_i, \min(on_i(C^*), nov(t_i, t))\rangle$, according to the new C^*.

In essence, for each cluster, the algorithm initially selects the triple that simultaneously maximizes *document frequency* in the news corpus and *Inner Novelty* from the set of viable candidates. Subsequently, it iterates the selection process. At each iteration, a newly selected triple must itself exhibit high relevance and *Inner Novelty*, while also demonstrating substantial *Outer Novelty* concerning the previously selected triples.

Keeping in mind the $C1, C2, C3$ clusters from the example described earlier, the outcome of the best triples selection process leads to the following situation: $C1$ remains intact, while in $C2$, redundant triples like ("*Covid*", "*infect*", "*patients*") and ("*Hospital*", "*treat*", "*cure*") are pruned. Afterwards the user can actively engage in the analysis by potentially removing triples, adding new ones to better contextualize each cluster's analysis, deleting clusters or even suggesting additional clusters for useful facets of the analysis overlooked by the system. Let's assume the user eliminates $C3$ as it is of little interest for user analysis purposes.

The final step involves transforming the identified SVO triples into prompts suitable for a Zero-shot classification system, achieved through the utilization of a Large Language Model (LLM) (Touvron et al., 2023; Jiang et al., 2023; OpenAI, 2023). For each cluster, all selected triples are fed into an LLM alongside a prompt designed to convert them into assertive forms that can define concepts. This process may involve synthesizing and aptly combining the contributions of all triples within the cluster. An important factor is the variable number of triples per cluster. However, the LLM can also be tasked with generating m distinct assertions, accommodating the diversity and breadth of information captured by the cluster's triples. This step of **Prompt Generation** is implemented by making a request to the LLM through prompts such as: "*Consider the input triples consisting of 3 terms. I need you to generate 3 sentences, where each sentence serves to 'describe' the triples. The sentences you generate must follow the format 'This news is about <complete>' and MUST NOT exceed 12 words in length. The triples are as follows:*". After analyzing LLMs such as LLAMA (Touvron et al., 2023), Mistral (Jiang et al., 2023), and GPT-4 (OpenAI, 2023), we have chosen to employ GPT-4 for the quality of the generated prompts. Currently, our selection is solely based on empirical analyses to determine the "best" LLM: a more comprehensive examination is still underway. For instance, using the above triples from the example cluster C_2, statements such as "*This news is about hospitals admitting patients during the pandemic.*" or "*This news is about hospitals vaccinating patients to combat illnesses.*" or "*This news is about Covid infecting lungs, posing respiratory risks.*" are generated. A similar strategy is applied to provide a title for the cluster. The request "*Write a name that precisely describes the following set of word triples. Please respond with ONLY ONE name consisting of ONLY 2 or 3 WORDS, the triples are as follows:*" is adopted. Taking the example further, the names of the following emerging topics are generated starting, respectively, from the sets $C1, C2$: "COVID INFECTIONS", "COVID HEALTHCARE DYNAMICS".

3.3. Zero-shot Classification

In the final step, **Zero-shot Classification** is applied to match the amassed news articles with the identified topics, significantly enhancing the article metadata granularity associated with specific thematic facets. Inspired by (Yin et al., 2019), our approach employs a Zero-shot classifier built upon Large Language Models (LLMs) that requires no fine-tuning. This method aligns articles to topics by treating the task as one of natural language inference (NLI), following paradigms established in (Dagan et al., 2013; Bowman et al., 2015). Specifically, it assigns each article (treated as the premise in a classical textual entailment task) to a topic based on the degree to which the LLM deems the text to logically infer the topic's defining prompt (treated as the hypothesis corresponding to the premise). This process ensures articles are categorically aligned with topics through a logical inference mechanism, offering a precise and context-aware topic association. In our approach, we have employed a model based on BART (Lewis et al., 2020), trained on the Multi-Genre Natural Language Inference (MultiNLI) corpus (Williams et al., 2018), a crowd-sourced collection of 433k sentence pairs annotated with textual entailment information across various genres, including news[3]. The process involves presenting a news article and, in turn, each prompt of each generated topic, calculating a score, in terms of probability of the "truth" of the entailment. This evaluation allows us to determine which prompts are activated by the news article along with the corresponding probability.

For instance, let's consider a scenario where we

[3] https://huggingface.co/facebook/bart-large-mnli

have a news item: "*Rome, Italy - As the Eternal City faces a concerning rise in COVID-19 cases, Rome's hospitals are stepping up their response with an aggressive vaccination campaign. The regional health authorities have reported a significant increase in hospital admissions due to COVID-19, prompting a swift reaction from the medical community. ...*" each prompt related to the overarching theme of 'COVID HEALTHCARE DYNAMICS" is coupled with the original news excerpt to form a composite input for the Zero-shot classifier. For instance, the news snippet "*Rome, Italy - As the Eternal City faces a concerning rise in COVID-19 cases ...*" is appended with a separator `[SEP]` followed by a prompt such as "*This news is about hospitals vaccinating patients to combat illnesses.*" This methodology allows the classifier to generate an internal representation of the combined input and evaluate it against predefined categories that determine whether the second statement is implied by the first. For each pairing, the classifier assigns an entailment probability score, reflecting the relevance of the prompt to the original text within the context of the selected theme. From the examples provided, the prompt stating "*This news is about hospitals vaccinating patients to combat illnesses.*" yielded an entailment score of 0.87, indicating a strong connection to the topic at hand. Conversely, the statement "*This news is about hospitals vaccinating patients to combat illnesses.*" received a lower score of 0.68, suggesting it is related but does not capture the core aspects of the news item as effectively. The prompt "*This news is about Covid infecting lungs, posing respiratory risks.*" regarding COVID-19 affecting the lungs with a score of 0.05 was not triggered, likely due to the absence of direct reference to lung infections in the news piece. From this small sample, it becomes clear that the prompt with a score of 0.87 is identified as a significant trigger for the corresponding theme cluster, effectively encapsulating the primary focus of the news item. The prompt with a score of 0.68, while relevant, may not fully capture the salient aspects of the scenario. In contrast, the prompt with a score of 0.01 is deemed unrelated, highlighting the classifier's ability to discern relevance based on the specificity of the information provided about the lungs.

Subsequently, an overall probability for the topic itself is based on a composition of each of those derived from the m individual prompts. Given m possible prompts and their corresponding entailment probabilities, we investigated the following policies: **Maximum Probability**: Assigns the maximum score observed across prompts to the topic. This greedy approach, however, is exposed to imprecise outliers or spikes; **Top-Best Average**: Returns the average score from the top b scores among the topic's prompts. This method is more robust against outliers; **Topic Average**: Computes the average score across all m prompts. The topic is activated only when, on average, the news article verifies all aspects described by the prompts.

Considering two topic clusters with different sets of entailment scores for their prompts—0.87, 0.68, 0.05 for the first cluster, and 0.88, 0.19, 0.02 for a second hypothetical cluster—we evaluate topic assignment policies. The Maximum Probability method, focusing on the highest score per cluster, initially suggests the second cluster (0.88) as more relevant than the first (0.87). Yet, this overlooks the composite thematic relevance of all prompts. Adopting the Top-Best Average method with $b = 2$, we find that the average of the top two scores offers a more nuanced perspective: 0.775 for the first cluster versus 0.535 for the second. This method, by mitigating outlier influence, indicates a stronger alignment of the first cluster with the news, demonstrating the importance of a broader assessment beyond single peak scores for more accurate topic relevance evaluation.

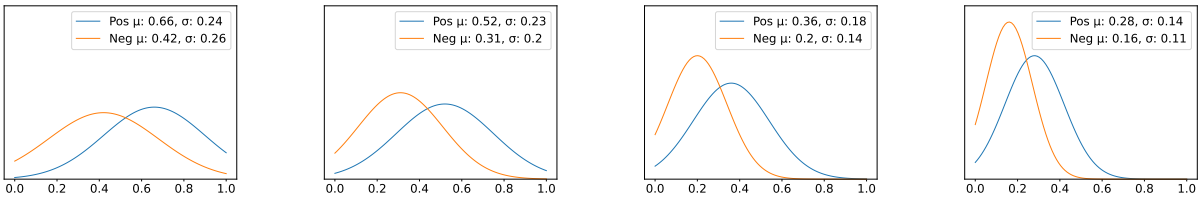
4. Experimental Evaluation

The experimental section is designed to evaluate the effectiveness of our workflow, which includes a series of complex steps ranging from web page collection, and topic generation, to the 0-shot classification of retrieved news articles. The core hypothesis of this experiment is that the model is effective if it can accurately re-process news articles previously categorized by the analysts under specific themes. It should regenerate consistent topics and associate the news articles to the topics coherently with the analysts' original choices. The workflow's success is measured according to the comparison between probabilities assigned to the method and those expected, as derived from the analyst annotation. First, the model is used to generate distinct topics based on the input analyst classes: in this way, the initial association between news articles and their original theme is known. Subsequently, if the system classifies the news articles according to the newly generated topics, and these topics align closely with the input themes, then the model is performing in harmony with the analyst's expectation.

Experimental setup. During the surveillance period from February 2020 to September 2022, analysts concentrated on monitoring COVID-19 outbreaks across various epidemiological settings. Within this timeframe, ISS experts manually categorized a total of 2,254 news articles, associated to "COVID VARIANTS" (313 news), "NURSING HOMES OUTBREAKS" (682), "HOSPITAL OUTBREAKS" (417), "SCHOOL OUTBREAKS" (574) and FAMILY/FRIEND

OUTBREAKS" (268). It's important to note that for their analysis, the analysts focused on a subset of topics at a time and could only associate a news article with a subset of those that manifested, such as a news piece discussing an outbreak in a hospital and then in a nursing home. To generate the topics and attempt to replicate the analysis performed on the documents, we selected a couple of seed words for each input theme: "COVID VARIANTS" corresponds to "*variant*" and "*English variant*", "*omicron variant*", "*delta variant*"; "NURSING HOMES" is represented by "*healthcare worker*" and "*elderly*", "*healthcare residence*"; for "HOSPITAL OUTBREAKS" the seeds are "*Hospital*", "*department*", "*patient*", "*contagion*"; "SCHOOL OUTBREAKS" is expressed by "*school*", "*remote learning*", "*student*", "*teacher*"; and "FAMILY/FRIEND OUTBREAKS" corresponds to "*parent*", "*family member*", "*condominium*", "*relative*". To generate the topics, seed words were selected for each input theme, and parameters were evaluated to model the process closely. For the linguistic triple generation step, we expanded the search to include a broad range of verbs and nouns, specifically setting $e_v = 150$ for verbs and $e_n = 100$ for nouns. The process was carried out for each topic individually, applying Triple Clustering with values of $k = 2, 3, 4$. This range was chosen to prevent an overwhelming proliferation of clusters while still capturing a diverse array of topics. To mimic the analytical phase typically performed by an analyst, clusters were selected based on which k values yielded the most coherent and relevant topics. The pruning process, as outlined in Algorithm 1, was guided by specific parameters for measuring novelty within and between the generated triples. For *Inner Novelty*, the parameters were set as $\beta^{\mathcal{SV}} = 0.25$, $\beta^{\mathcal{SO}} = 0.65$, and $\beta^{\mathcal{VO}} = 0.10$. These parameters helped assess how distinct each SVO triple argument was from the others, ensuring a richer semantic variety within the topics. *Outer Novelty*, which measures the diversity between triples, was regulated with $\gamma^{\mathcal{S}} = 0.25$, $\gamma^{\mathcal{O}} = 0.10$, and $\gamma^{\mathcal{V}} = 0.65$. The threshold $\epsilon = 0.3$ in the algorithm ensured that only the triples significantly different from those already selected were chosen, simulating an analyst's decision-making process in refining the topics. Ultimately, this approach led to the generation of 11 distinct clusters for 44 total prompts. For each of these generated topics, the system requested m prompts to facilitate the prompt generation phase. The detailed list of clusters and their respective prompts, which form the backbone of our topic generation and classification process, is provided in Appendix A. Finally, the individual news articles were associated with topics by applying the Zero-shot classifier, determining an association score for each of the 44 prompts.

Results. To assess the quality of the classification system in accurately associating news articles with the correct topics, aligning with the input themes, we examined the system's ability to assign a score to each news article for the single class identified by the analyst (referred to as the positive class) used to distinguish it through the probabilities associated with unrelated themes (referred to as negative classes). Notice that, as we cannot rely on the hypothesis that each news article belongs to only one correct class, studying the overall behaviors of the method requires studying its probability distributions. Probability scores depend on the three policies defined for Zero-shot classification. Another reference policy can be added, allowing a news article to be associated with the average value of all prompts generated from the seeds related to that input topic or associated class, thus called the **Class average** policy. In Figure 1, the distributions of scores associated with each generated topic are shown for each defined policy. The distributions of scores for positive (pos) and negative (neg) classes are depicted in blue and orange respectively, by assuming a normal distribution given by its mean (μ) and standard deviation (σ). It is interesting to note how, for each policy, the system is shown capable of separating the distributions, confirming its ability to re-associate news articles with topics consistently with the original analysts' classifications. Clearly, the smaller the intersection between the two curves, the greater the system's ability to replicate the work of the analysts. It is interesting to note that in the policy called Best Probability, selecting the prompt that maximizes the association probability by observing only one prompt in the class has a high mean for positives ($\mu = 0.66$ in Figure 1 (a)). Unfortunately, there are also several spikes for the discarded classes with a mean ($\mu = 0.42$). This could be due to other topics, generated for other input themes, being often activated, apparently in disagreement with the analyst. However, generally, analysts have only selected one major topic, and it is possible they discarded other topics that are part of the article discussion. Moving to Figure 1 (b), by taking the average of the first $b = 2$ in the Top-Best Average policy, the average values predictably tend to decrease, and interestingly, the standard deviation also decreases because the spikes are mitigated. This phenomenon is further evident in Figure 1 (c) where averaging all prompts in a generated topic forces a topic to cover practically all sub-themes discussed in the topic, according to the Topic Average policy. Pushing the situation further in Figure 1 (d), where all prompts of all topics generated for a class are averaged, it is evident that the means significantly decrease (dropping to $\mu = 0.28$ for positives) but the negatives are practically nullified (with a $\mu = 0.16$) and the standard deviations are greatly reduced, also reducing the



(a) Best probability (b) Top-Best Average (c) Topic Average (d) Class Average

Figure 1: Distributions of scores for generated topics per policy, showing positive and negative classes with the corresponding means (μ) and standard deviations (σ).

intersection area between the curves. This suggests the importance of having multiple prompts in individual topics to make the system more stable and avoid associations due to possible spikes.

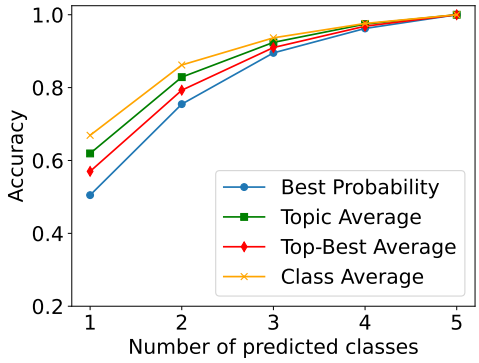


Figure 2: Accuracy

We then assessed the system's ability to reclassify individual news articles, ranking the classes based on the scores suggested by the system's topics, as shown in Figure 2: the y-axis represents accuracy, while the x-axis lists the classes. As mentioned, it's not necessarily the case that if the first class proposed by the system is incorrect, the system's handling is flawed. Therefore, we calculated accuracy for correctly identifying just the first class (c=1), but also for c=2, then c=3, up to c=5 (where, by construction, accuracy is 1.0). The Best Probability policy confirms its vulnerability to spikes, with accuracy at c=1 of 0.51 and c=2 of 0.75. This is significantly compensated by the various policies, as evidenced by the distributions in Figure 1, with Top-best Avg at c=1 having an accuracy of 0.57, Topic Average at c=1 rising to 0.62, and then Class Average at c=1 reaching 0.67. Interestingly, more than 88% of the news articles are reassociated with the analysts' topics if only the first two system suggestions are considered under the Class Average policy. The experimental results emphasize the benefits of our approach, in accurately mapping news articles to relevant topics, demonstrating the workflow's potential to streamline Epidemic Intelligence processes. An error analysis is reported in Appendix B.

5. Conclusion

This study presented a novel framework utilizing Large Language Models (LLMs) to enhance Epidemic Intelligence through automated topic discovery and 0-shot classification. We aimed to address in this way the challenge of effectively identifying and categorizing potential health hazards, with a focus on the COVID-19 pandemic. Our methodology diverges from traditional probabilistic models by offering explicit analytical support through the generation of actionable topic statements, thereby facilitating a Zero-shot classification mechanism that accurately matches news articles to defined topics without resorting to fine-tuning. Our methodology, integrating a decoder LLM, faces potential limitations highlighted by (Huang et al., 2023), such as susceptibility to hallucinations. This affects the generation of prompts and topic names, which could lead to inaccurate descriptions of emerging arguments. Our approach, however, is not designed as an inflexible, fully automated system that diminishes the role of analysts. Rather, it is intended to enhance and enrich their analytical capabilities, promoting an interactive and collaborative exploration of data. Analysts maintain the capacity to modify outputs at any stage, enabling them to select relevant topics or refine prompts for Zero-shot classification, thereby ensuring a more accurate and insightful analysis.

The results from our experimental evaluation highlight the robustness and effectiveness of our workflow in aligning with the analytic processes traditionally employed by experts in Epidemic Intelligence. The implementation of multiple classification policies has demonstrated a significant improvement in the system's ability to accurately associate news articles with relevant topics. This advancement is evident in the increased accuracy rates and the reduction of classification errors, underscoring the system's capacity to handle complex thematic categorizations reliably.

Acknowledgements

The project realized with the technical and financial support of the Ministry of Health - CCM.

Bibliographical References

Aly Abdelrazek, Yomna Eid, Eman Gawish, Walaa Medhat, and Ahmed Hassan. 2023. Topic modeling algorithms and applications: A survey. *Information Systems*, 112:102131.

Stanislaw Adaszewski, Pascal Kuner, and Ralf J Jaeger. 2021. Automatic pharma news categorization. *arXiv preprint arXiv:2201.00688*.

Mohammed Ali Al-Garadi, Yuan-Chi Yang, and Abeed Sarker. 2022. The role of natural language processing during the covid-19 pandemic: Health applications, opportunities, and challenges. *Healthcare*, 10(11).

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press.

David M Blei and John D Lafferty. 2009. Topic models. In *Text mining*, pages 101–124. Chapman and Hall/CRC.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Zied Bouraoui, José Camacho-Collados, and Steven Schockaert. 2020. Inducing relational knowledge from BERT. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7456–7463. AAAI Press.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.

Kevin Matthe Caramancion. 2023. News verifiers showdown: a comparative performance evaluation of chatgpt 3.5, chatgpt 4.0, bing ai, and bard in news fact-checking. *arXiv preprint arXiv:2306.17176*.

Ambrish Choudhary, Mamatha Alugubelly, and Rupal Bhargava. 2023. A comparative study on transformer-based news summarization. In *2023 15th International Conference on Developments in eSystems Engineering (DeSE)*, pages 256–261. IEEE.

Rob Churchill and Lisa Singh. 2022. The evolution of topic modeling. *ACM Computing Surveys*, 54(10s):1–35.

Nigel Collier, Son Doan, Ai Kawazoe, Reiko Matsuda Goodwin, Mike Conway, Yoshio Tateno, Quoc-Hung Ngo, Dinh Dien, Asanee Kawtrakul, Koichi Takeuchi, et al. 2008. Biocaster: detecting public health rumors with a web-based text mining system. *Bioinformatics*, 24(24):2940–2941.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. *Recognizing Textual Entailment: Models and Applications*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Martina Del Manso, Daniele Petrone, Matteo Spuri, Chiara Sacco, Alberto Mateo Urdiales, Roberto Croci, Stefania Giannitelli, Patrizio Pezzotti, Daniele Mipatrini, Francesco Maraglino, et al. 2022. Il sistema di sorveglianza basato su eventi in italia dal 2009 al 2021: verso una intelligence di sanità pubblica. *Bollettino epidemiologico nazionale*.

Lin Deping, Wang Hongjuan, Liu Mengyang, and Li Pei. 2021. News text classification based on bidirectional encoder representation from transformers. In *2021 International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA)*, pages 137–140. IEEE.

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jerome Friedman. 2002. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38:367–378.

Jerome H. Friedman. 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189 – 1232.

Anthony Gillioz, Jacky Casas, Elena Mugellini, and Omar Abou Khaled. 2020. Overview of the transformer-based models for nlp tasks. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, pages 179–183. IEEE.

Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2013. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524.

Yoav Goldberg and Omer Levy. 2014. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method.

Ramanathan Guha, Rob McCool, and Eric Miller. 2003. Semantic search. In *Proceedings of the 12th international conference on World Wide Web*, pages 700–709. ACM.

Erkan Gunes and Christoffer Koch Florczak. 2023. Multiclass classification of policy documents with large language models. *arXiv preprint arXiv:2310.08167*.

Anushka Gupta, Diksha Chugh, Anjum, and Rahul Katarya. 2022. Automated news summarization using transformers. In *Sustainable Advanced Computing: Select Proceedings of ICSAC 2021*, pages 249–259. Springer.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The elements of statistical learning: data mining, inference and prediction*, 2 edition. Springer.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

Nancy Ide and Jean Véronis. 1990. Very large neural networks for word sense disambiguation. In *9th European Conference on Artificial Intelligence, ECAI 1990, Stockholm, Sweden, 1990*, pages 366–368.

Network Intelligence and Alessandro Miglietta. 2022. Istituto superiore di sanitá il sistema di sorveglianza basato su eventi in italia dal 2009 al 2021: verso una intelligence di sanitá pubblica. *Scientific reports of the Istituto superiore di sanitá*, pages 19–28.

Frederick Jelinek. 1998. *Statistical Methods for Speech Recognition (Language, Speech, and Communication)*. The MIT Press.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Daniel Jurafsky and James H. Martin. 2009. *Speech and language processing*, 2. ed., [pearson international edition] edition. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International, London [u.a.].

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788.

Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee, and Khairullah Khan. 2010. A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*, 1(1):4–20.

Saima Khosa, Arif Mehmood, and Muhammad Rizwan. 2023. Unifying sentence transformer embedding and softmax voting ensemble for accurate news category prediction. *Computers*, 12(7):137.

Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. 2019. Text classification algorithms: A survey. *Information*, 10(4):150.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, number 2 in Proceedings of Machine Learning Research, pages 1188–1196, Bejing, China. PMLR.

Michael Lebowitz. 1988. The use of memory in text processing. *Commun. ACM*, 31:1483–1502.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S Yu, and Lifang He. 2022. A survey on text classification: From traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(2):1–41.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2022. A survey of transformers. *AI Open*, 3:111–132.

Tie-Yan Liu. 2009. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331.

Yuandong Luan and Shaofu Lin. 2019. Research on text classification based on cnn and lstm. In *2019 IEEE international conference on artificial intelligence and computer applications (ICAICA)*, pages 352–355. IEEE.

G. Madhu, Dr. A. Govardhan, and Dr. T. V. Rajinikanth. 2011. Intelligent semantic web search engines: A brief survey.

C.D. Manning, P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

Jon Mcauliffe and David Blei. 2007. Supervised topic models. *Advances in neural information processing systems*, 20.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning–based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, 54(3):1–40.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Jesse O'Shea. 2017. Digital disease detection: A systematic review of event-based internet biosurveillance systems. *International journal of medical informatics*, 101:15–22.

C Paquet, D Coulombier, R Kaiser, and M Ciotti. 2006. Epidemic intelligence: a new framework for strengthening disease surveillance in europe. *Eurosurveillance*, 11(12):5–6.

Loreto Parisi, Simone Francia, and Paolo Magnani. 2020. Umberto: an italian language model trained with whole word masking. https://github.com/musixmatchresearch/umberto.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, David Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. 2020. Hopfield networks is all you need. Cite arxiv:2008.02217Comment: 10 pages (+ appendix); 12 figures; Blog: https://ml-jku.github.io/hopfield-layers/; GitHub: https://github.com/ml-jku/hopfield-layers.

Shaina Raza and Brian Schwartz. 2023. Constructing a disease database and using natural language processing to capture and standardize free text clinical information. *Scientific Reports*, 13(1):8591.

Shaina Raza, Brian Schwartz, and Laura C Rosella. 2022. Coquad: a covid-19 question answering dataset system, facilitating research, benchmarking, and practice. *BMC bioinformatics*, 23(1):1–28.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Flavia Riccardo, Martina Del Manso, Maria Grazia Caporali, Christian Napoli, Jens P Linge, Eleonora Mantica, Marco Verile, Alessandra Piatti, Maria Grazia Pompa, Loredana Vellucci, Virgilio Costanzo, Anan Judina Bastiampillai, Eugenia Gabrielli, Maria Gramegna, and Silvia Declich. 2016. Event-Based surveillance during EXPO milan 2015: Rationale, tools, procedures, and initial results. *Health Secur*, 14(3):161–172.

Flavia Riccardo, Mika Shigematsu, Chow Catherine, Mcknight Jason, Jens Linge, Brian Doherty, Maria Dente, Silvia Declich, Barker Mike, Barboza Philippe, Laetitia Vaillant, Donachie Alastair, Mawudeku Abla, Blench Michael, and Arthur Ray. 2014. Interfacing a biosurveillance portal and an international network of institutional analysts to detect biological threats. *Biosecurity and bioterrorism: biodefense strategy, practice, and science*, 12:325–36.

Agnès Rortais, Jenya Belyaeva, Monica Gemo, Erik Van der Goot, and Jens P Linge. 2010.

Medisys: An early-warning system for the detection of (re-) emerging food-and feed-borne hazards. *Food Research International*, 43(5):1553–1556.

Magnus Sahlgren. 2006. *The Word-Space Model*. Ph.D. thesis, Stockholm University.

Isabel N Santana, Raphael S Oliveira, and Erick GS Nascimento. 2022. Text classification of news using transformer-based models for portuguese. *Journal of Systemics, Cybernetics and Informatics*, 20(5):33–59.

Souvika Sarkar, Dongji Feng, and Shubhra Kanti Karmaker Santu. 2023. Zero-shot multi-label topic inference with sentence encoders and llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16218–16233.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *International Conference on Acoustics, Speech and Signal Processing*, pages 5149–5152.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. *arXiv preprint arXiv:2305.08377*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. Cite arxiv:2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Yu Wang, Yuan Wang, Zhenwan Peng, Feifan Zhang, Luyao Zhou, and Fei Yang. 2023. Medical text classification based on the discriminative pre-training model and prompt-tuning. *Digital Health*, 9:20552076231193213.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Basil Blackwell, Oxford.

World Health Organization. 2008. Communicable disease alert and response for mass gatherings. In *Technical Workshop. Geneva, Switzerland*, pages 29–30.

World Health Organization. 2014. Early detection, assessment and response to acute public health events: implementation of early warning and response with a focus on event-based surveillance: interim version. Technical report, World Health Organization.

Roman Yangarber and Ralf Steinberger. 2009. Automatic epidemiological surveillance from on-line news in medisys and puls. In *IMED-2009: International Meeting on Emerging Diseases and Surveillance (2009)*.

K Yasaswi, Vijaya Kumar Kambala, P Sai Pavan, M Sreya, and V Jasmika. 2022. News classification using natural language processing. In *2022 3rd International Conference on Intelligent Engineering and Management (ICIEM)*, pages 63–67. IEEE.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. 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 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, and Eric Altendorf. 2016. Semi-supervised word sense disambiguation with neural models.

A. Emerging Topics and prompts.

In this section, we present the topics and prompts that were generated and used in the experimental evaluation detailed in Section 4. For each cluster, we list the original prompts used in the experimentation and, for the convenience of the reader, their translation into English.

A.1. Topic: Covid Variants

C1: "*Diffusione delle varianti*" / "*Spread of variants*"

- "*Questa notizia riguarda una variante che richiede un accurato sequenziamento per monitorarne l'evoluzione.*" / "*This news is about a variant that requires careful sequencing to monitor its evolution.*"
- "*Questa notizia riguarda la capacità di sequenziare le varianti in casi complessi.*" / "*This news is about the ability to sequence variants in complex cases.*"
- "*Questa notizia riguarda come le varianti possono mutare il virus nel tempo.*" / "*This news is about how variants can mutate the virus over time.*"
- "*Questa notizia riguarda la diffusione delle varianti.*" / "*This news is about the spread of variants.*"

C2: "*Diffusione variante inglese*" / "*Spread of the English variant*"

- "*Questa notizia riguarda la vaccinazione per non incorrere in aumento di casi positivi.*" / "*This news is about vaccination to avoid an increase in positive cases.*"
- "*Questa notizia riguarda la necessità di vaccinare la popolazione contro la variante inglese.*" / "*This news is about the need to vaccinate the population against the English variant.*"
- "*Questa notizia riguarda la scoperta di un caso positivo relativo alla variante inglese.*" / "*This news is about the discovery of a positive case related to the English variant.*"
- "*Questa notizia riguarda la diffusione della variante inglese.*" / "*This news is about the spread of the English variant.*"

A.2. Topic: Nursing Homes Outbreaks

C3: "*Cura degli ospiti nelle RSA*" / "*Care of residents in nursing homes*"

- "*Questa notizia riguarda la necessità di ricoverare pazienti in una residenza sanitaria per anziani (RSA).*" / "*This news is about the need to hospitalize patients in a nursing home for the elderly.*"
- "*Questa notizia riguarda le procedure necessarie per accogliere un ospite nella residenza sanitaria per anziani (RSA).*" / "*This news is about the necessary procedures to welcome a guest into the nursing home for the elderly.*"
- "*Questa notizia riguarda le misure prese per isolare un ospite infetto nella residenza sanitaria per anziani (RSA).*" / "*This news is about the measures taken to isolate an infected guest in the nursing home for the elderly.*"
- "*Questa notizia riguarda la cura degli ospiti nelle RSA.*" / "*This news is about caring for guests in nursing homes for the elderly.*"

C4: "*Covid negli ospizi*" / "*Covid in nursing homes*"

- "*Questa notizia riguarda l'importante compito di vaccinare gli anziani contro il COVID.*" / "*This news is about the important task of vaccinating the elderly against COVID.*"
- "*Questa notizia riguarda l'importanza di isolare le residenze sanitarie per anziani (RSA) per prevenire focolai di COVID.*" / "*This news is about the importance of isolating nursing homes for the elderly to prevent COVID outbreaks.*"
- "*Questa notizia riguarda la necessità di ricoverare i pazienti COVID nelle residenze sanitarie per anziani (RSA) per cure adeguate.*" / "*This news is about the need to hospitalize COVID patients in nursing homes for the elderly for proper care.*"
- "*Questa notizia riguarda il covid negli ospizi.*" / "*This news is about covid in nursing homes.*"

A.3. Topic: Hospital Outbreaks

C5: "*Gestione dell'emergenza in ospedale*" / "*Emergency management in the hospital*"

- "*Questa notizia riguarda il paziente che risultò positivo al test presso la struttura ospedaliera.*" / "*This news is about the patient who tested positive at the hospital facility.*"
- "*Questa notizia riguarda il reparto di terapia intensiva all'interno dell'ospedale.*" / "*This news is about the intensive care unit within the hospital.*"
- "*Questa notizia riguarda un reparto dell'ospedale pronto a soccorrere ogni paziente.*" / "*This news is about a hospital department ready to assist every patient.*"
- "*Questa notizia riguarda la gestione dell'emergenza in ospedale.*" / "*This news is about the emergency management in the hospital.*"

C6: "*Impatto dell'epidemia: Ricovero ospedaliero*" / "*Impact of the epidemic: Hospitalization*"

- "*Questa notizia riguarda un ospedale che offre servizi di ricovero per i residenti.*" / "*This news is about a hospital that provides hospitalization services for residents.*"

- "*Questa notizia riguarda l'ospedale che si occupa di ricoverare i casi di COVID.*" / "*This news is about the hospital that takes care of hospitalizing COVID cases.*"
- "*Questa notizia riguarda un ospedale in una città, dove vengono ricoverate persone malate.*" / "*This news is about a hospital in a city where sick people are hospitalized.*"
- "*Questa notizia riguarda il ricovero ospedaliero dei casi.*" / "*This news is about the hospitalization of cases.*"

C7: "*Impatto dell'epidemia: Contagio in ospedale*" / "*Impact of the epidemic: Hospital contagion*"

- "*Questa notizia riguarda un paziente che risulta positivo al virus in ospedale.*" / "*This news is about a patient who tested positive for the virus in the hospital.*"
- "*Questa notizia riguarda il paziente che è risultato negativo al test.*" / "*This news is about the patient who tested negative.*"
- "*Questa notizia riguarda il contagio in ospedale che si può prevenire vaccinando ogni caso.*" / "*This news is about hospital contagion that can be prevented by vaccinating each case.*"
- "*Questa notizia riguarda i contagi che avvengono in ospedale.*" / "*This news is about the contagions that occur in the hospital.*"

A.4. Topic: School Outbreak

C8: "*Impatto della didattica a distanza*" / "*Impact of distance learning*"

- "*Questa notizia riguarda la chiusura della scuola conseguente all'attivazione della didattica a distanza.*" / "*This news is about the school closure following the activation of distance learning.*"
- "*Questa notizia riguarda l'importanza di rispettare le misure di sicurezza a scuola.*" / "*This news is about the importance of respecting safety measures at school.*"
- "*Questa notizia riguarda gli studenti che frequentano le scuole durante l'epidemia.*" / "*This news is about the students attending schools during the epidemic.*"
- "*Questa notizia riguarda l'impatto della didattica a distanza sulla scuola.*" / "*This news is about the impact of distance learning on school.*"

C9: "*Conseguenze dell'epidemia nella scuola*" / "*Consequences of the epidemic in schools*"

- "*Questa notizia riguarda l'insegnante che ha contribuito a contagiare un focolaio a scuola.*" / "*This news is about the teacher who contributed to infecting a outbreak at school.*"
- "*Questa notizia riguarda lo studente che risulta detenere il virus positivo.*" / "*This news is about the student who tested positive for the virus.*"
- "*Questa notizia riguarda le conseguenze dell'epidemia nella scuola.*" / "*This news is about the consequences of the epidemic in schools.*"

A.5. Topic: Family Outbreaks

C10: "*Impatto familiare della pandemia*" / "*Family impact of the pandemic*"

- "*Questa notizia riguarda un familiare che è stato contagiato in un focolaio.*" / "*This news is about a family member who was infected in an outbreak.*"
- "*Questa notizia riguarda un genitore che è stato vaccinato ma risulta positivo.*" / "*This news is about a parent who has been vaccinated but tested positive.*"
- "*Questa notizia riguarda un focolaio familiare che ha contagiato molte persone.*" / "*This news is about a family outbreak that has infected many people.*"
- "*Questa notizia riguarda l'impatto sulle famiglie della pandemia.*" / "*This news is about the impact on families of the pandemic.*"

C11: "*Vaccinazione familiare COVID*" / "*Family COVID vaccination*"

- "*Questa notizia riguarda la vaccinazione di gruppi di persone conoscenti.*" / "*This news is about the vaccination of groups of acquaintances.*"
- "*Questa notizia riguarda il timore di un familiare di contagiare altre persone con il coronavirus.*" / "*This news is about the fear of a family member of infecting other people with the coronavirus.*"
- "*Questa notizia riguarda l'importanza di vaccinare per proteggere la salute di parenti e amici.*" / "*This news is about the importance of vaccinating to protect the health of relatives and friends.*"
- "*Questa notizia riguarda la vaccinazione di famiglie contro il COVID.*" / "*This news is about the vaccination of families against COVID.*"

B. Error analysis

In the experimental evaluation reported in Section 4, the system manages to classify 67% of the news articles when a single class is proposed, and this figure rises to more than 88% if only the first two system suggestions are considered under the Class Average policy.

To understand the reason behind the discrepancy in these results, we examined the confusion matrix presented in Figure 3, which shows which classes were confused with each other. In the matrix, the analysts' annotations are on the rows, and the system's proposed associations when only one class is proposed are on the columns. Most of the classifications considered incorrect when only the first class is proposed (but corrected when two are proposed, indicating the second is correct) are news articles originally associated with the input topic Nursing Homes Outbreaks or School Outbreaks, which are classified as Family/Friend Outbreaks, or Nursing Homes Outbreaks classified as Hospital Outbreaks.

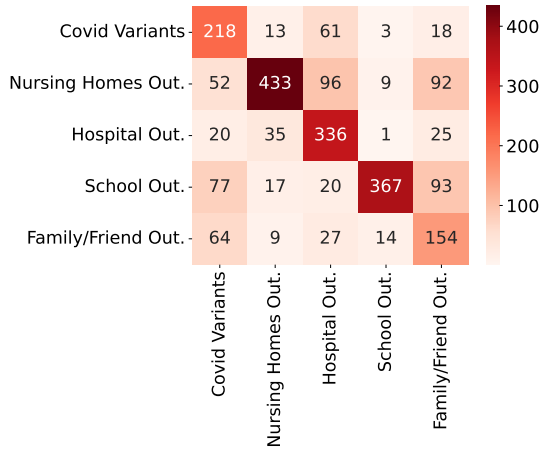


Figure 3: Confusion matrix

This observation implies that discussions about nursing homes may naturally reference families rather than hospitals, as these topics are inherently related. Motivated by this, we conducted a manual analysis of cases that would have been deemed errors. This deeper examination may reveal nuances in the data that automated classification initially overlooked, underscoring the complex interplay between seemingly distinct topics and the importance of contextual understanding in accurately categorizing news articles.

For example, consider the following news article:

> "*«La variante inglese era presente in un caso su cinque, ma nelle ultime due settimane la diffusione è molto aumentata». Così parlava tre giorni fa la dottoressa Antonia Ricci, direttrice generale dell'istituto zooprofilattico di Legnaro. Ora gli effetti si vedono e fanno paura: un focolaio è stato registrato in un gruppo di bambini di San Martino di Lupari e poi il contagio si è propagato in diverse zone dell'Alta Padovana costringendo l'Ulss ad una decisione drastica: quattro scuole chiuse. Il sindaco è stato invitato ad avviare la didattica a distanza per l'asilo Campagnalta e per la scuola elementare Sauro, dove è stato registrato il primo cluster. Stesso provvedimento anche per la materna Almarech di Villa del Conte e soprattutto per il liceo Tito Lucrezio Caro di Cittadella.*"

The system predicted the class as COVID VARIANTS, primarily due to the strong activation of the prompt "*This news is about the discovery of a positive case related to the English variant*" which received a confidence score of 0.98. The sentence "*The prevalence of the English variant has significantly increased in the past two weeks, transitioning...*", would suggest a specific relevance of the mentioned prompt to the news at hand. Conversely, the original classification was SCHOOL OUTBREAK, a category significantly represented by the prompt "*This news is about the school closure following the activation of distance learning*", achieving a confidence score of 0.84. In this case, it is worth noting how the prompt's score can be justified by the statement "*The mayor has been invited to initiate distance learning for the Campagnalta kindergarten and Sauro elementary school...*". Therefore, the system's prediction of COVID VARIANTS as the primary class, with a higher confidence score for the prompt related to the English variant, reflects the significant mention of the variant in the article. However, the original class SCHOOL OUTBREAK is also strongly represented, especially given the specific actions taken in response to the outbreaks in schools. This discrepancy suggests that while the variant's presence is a crucial aspect of the news, the article's core subject revolves around the implications of this presence on local schools. This case exemplifies the nuanced understanding required in classifying news articles, where multiple relevant themes can coexist, emphasizing the importance of considering all potential topics when classifying complex news stories.

Let us consider another example:

> "*Coronavirus. Ortona. Focolaio nella casa di riposo 'Don Bosco': 43 positivi. Contagi in ospedale nella struttura, che occupa i locali dell'ex Istituto salesiano, sono stati riscontrati 43 casi di Covid-19, dopo che sono stati effettuati tamponi a tappeto. Gli ospiti contagiati sono 33 e sono 10 coloro che, tra addetti e personale, hanno contratto l'infezione. "Situazione costantemente monitorata dalla Asl e da noi", dice il sindaco Leo Castiglione. Sette degli anziani positivi, quelli che presentano sintomi, sono stati trasferiti in ospedale a Chieti. Attenzione anche sull'ospedale "Bernabeo", dove il reparto di Lungodegenza si è trasformato in un focolaio, con 10 pazienti positivi. Erano otto ma nelle ultime ore i casi sono aumentati. Nel reparto al momento stop a ricoveri e a dimissioni. Nel Centro di procreazione medicalmente assistita, invece, sono cinque gli operatori sanitari che hanno preso il virus.*"

The article led to the system predicting HOSPITAL OUTBREAKS as the primary class, significantly influenced by the activation of the prompt "*This news is about the hospital that takes care of hospitalizing COVID cases.*" which received a high confidence score of 0.96. This prediction underscores the focus on hospital-related aspects of the outbreak, particularly the transfer of symptomatic elderly patients to a hospital and the emergence of a cluster

within the hospital's long-term care department. As is evident, for instance, in the sentence "*Seven of the elderly individuals who tested positive, those exhibiting symptoms, have been transferred to the hospital in Chieti*". However, the original classification was NURSING HOMES OUTBREAKS, which also finds strong representation through the activation of the prompt "*This news is about covid in nursing homes.*" with a confidence score of 0.89. This classification captures the article's primary focus on a COVID-19 outbreak in a nursing home, including the infection of residents and staff, which constitutes the core event described.

Finally, let us consider:

> "*Ladispoli, coronavirus nuovo focolaio alla Rsa Gonzaga. Parte il primo drive in della zona Finora 13 positivi nella struttura, tra degenti e operatori. Altro focolaio dopo una festa tra bambini: una mamma aveva il virus. Tre operai contagiati anche all'Fca di Cassino.*"

In this instance, the system classified the news article under FAMILY OUTBREAKS, predominantly due to the prompt "*This news is about a family member who was infected in an outbreak.*" being highly activated with a confidence score of 0.83. This classification highlights the mention of a family-related outbreak following a children's party within the article, which could explain the system's inclination towards the FAMILY OUTBREAKS class.

However, the intended classification was NURSING HOMES OUTBREAKS, which is significantly less represented in the system's evaluation, demonstrated by the most activated prompt, "*This news is about caring for guests in nursing homes for the elderly.*" receiving a lower confidence score of 0.34. This outcome indicates that while the article does mention a new outbreak at a nursing home ("*Rsa Gonzaga*") and provides a count of infected individuals, the mention of a family-related incident might have skewed the system's prioritization towards FAMILY OUTBREAKS.

Towards quantifying politicization in foreign aid project reports

Sidi Wang[♣♠], Gustav Eggers[‡♠], Alexia de Roode Torres Georgiadis[‡♠], Tuan Anh Đo[♢], Léa Gontard[♢], Ruth Carlitz[‡♠] and Jelke Bloem[♢♠]

♠ Data Science Centre, University of Amsterdam
♣ Computational Linguistics and Text Mining Lab, Vrije Universiteit Amsterdam
♢ Institute for Logic, Language and Computation, University of Amsterdam
‡ Department of Political Science, University of Amsterdam
s.wang4@student.vu.nl, gustav.eggers2@student.uva.nl,
alexia.de.roode.torres.georgiadis@student.uva.nl, tuan.do@student.uva.nl,
lea.gontard@student.uva.nl, r.d.carlitz@uva.nl, j.bloem@uva.nl

Abstract

We aim to develop a metric of politicization by investigating whether this concept can be operationalized computationally using document embeddings. We are interested in measuring the extent to which foreign aid is politicized. Textual reports of foreign aid projects are often made available by donor governments, but these are large and unstructured. By embedding them in vector space, we can compute similarities between sets of known politicized keywords and the foreign aid reports. We present a pilot study where we apply this metric to USAID reports.

Keywords: Politicization, foreign aid, document embeddings, Development Experience Clearinghouse

1. Introduction

When foreign aid is provided for political vs. altruistic interests, aid effectiveness is expected to suffer. However, evidence for this relationship – and the mechanisms through which it operates – is limited. This is due in large part to the fact that politicization tends to be operationalized quite bluntly. In addition, most studies of aid project effectiveness exclude the world's largest donor (the United States Agency for International Development, USAID), since USAID does not rate project effectiveness on a common numerical scale. However, the agency does make project evaluations publicly available through the agency's Development Experience Clearinghouse (DEC).[1]

The DEC provides access to over 10,000 evaluations spanning a range of activities and time periods. Unlike many of its peer foreign aid agencies, USAID does not have an independent evaluation agency but rather contracts evaluation out to various private firms. The evaluations thus comprise a range of formats and styles.

As a survey by Németh (2023) shows, NLP methods have been applied extensively and fruitfully to study the related notion of political polarization, showing that this concept can be successfully modeled on the basis of models trained on natural language data such as word embeddings. Unstructured natural language data is available in the DEC, annotated with categorical metadata representing variables of interest such as sectors. As the reports are fairly substantial (about 16k tokens per report on average) there should be enough in-domain training material for statistical NLP methods in these reports.

In this work, we aim to develop a metric of politicization by investigating whether this concept can be operationalized computationally. We also present a pilot study using a Doc2Vec-based method to quantify politicization of foreign aid reports in a sample of the DEC corpus.

2. Related work

In the context of foreign aid, politicization has occurred when "disagreements over the means to achieve a given goal are drawn along ideological lines that correspond to distinct political constituencies" (Carlitz, 2023, p. 9). This may affect the effectiveness of aid projects. The politicization of foreign aid has been studied primarily in terms of donor characteristics, as well as the dyadic relationships between particular donors and recipients. The most prominent operationalization of politicization considers whether donors and recipients are in some way aligned – where allegiances are measured using voting patterns in the UN general assembly (Bobba and Powell, 2007) or looking at joint membership in the UN Security Council (Dreher et al., 2018). Scholars have also examined the influence of political misalignment and ideological distance between donor and recipient governments (Dreher et al., 2015). Scholars have further inferred donor motives (and thus politicization) by examining the effect of aid given for developmental vs. 'strategic' purposes (Kilby and Dreher, 2010). Such blunt operationalizations make it diffi-

[1] https://dec.usaid.gov/dec/home/Default.aspx

Proceedings of the Second Workshop on Natural Language Processing for Political Sciences (PoliticalNLP 2024), pages 85–90

cult to distinguish relative politicization of different activities funded by the same donor, or otherwise provide for nuanced analysis.

2.1. Political NLP

The use of natural language processing to extract information from political texts and discourses has been explored from various angles, often driven by practical research questions. For example, one line of work is applying dimensionality reduction techniques, such as Latent Semantic Indexing (LSI), to identify political preferences in US elections (Bonica, 2013, 2014). Rheault and Cochrane (2019) investigated the potential of applying n-gram language modelling and Principal Component Analysis (PCA) for capturing ideological placements of parties in the US House.

Parallel efforts at the document level have employed NLP to analyze polarization in parliamentary systems (Peterson and Spirling, 2018), party affiliation (Yu et al., 2008) and news coverage (Chinn et al., 2020). Work on uncovering linguistic indicators of polarization often employs unsupervised learning methodologies. Moreover, the task of classifying political affiliations based on speech (Binder, 1999) and tweet texts (Demszky et al., 2019) has been explored with various machine learning algorithms, such as random forest classifiers. In the context of legal texts, Nay (2016) extended the Word2Vec model to embed institution-specific representations into a shared vector space, taking temporal relationships between them into account. This allows for the comparison of policy differences across US Congresses and sitting Presidents. However, in the landscape of international development projects, as Moore et al. (2023) note, there is a lack of work that specifically employs embedding techniques to extract, label and rate text from foreign aid evaluation reports.

Document embeddings have gained significant attention in the field of computational social science due to their ability of capturing abstract semantic information from textual data. Introduced by Le and Mikolov (2014), Doc2Vec extends the Word2Vec model (Mikolov et al., 2013) to generate a fixed-length representation of a given variable-length piece of text, allowing the model to be easily adapted to infer dense vector representations of sentences, paragraphs or entire documents in an unsupervised manner. There are two main approaches in Doc2Vec, so-called Distributed Bag-of-Words (DBOW) and Distributed Memory Paragraph Vectors (DMPV). DBOW treats each document as a single representation for context word prediction, ignoring the order of words within the document. DMPV preserves the order by using both document representation vector and the word vectors in the context to make predictions. Recent applications of Doc2Vec include sentiment analysis (Chen and Sokolova, 2021; Shuai et al., 2018; Liang et al., 2020), text classification (Dogru et al., 2021; Aubaid and Mishra, 2020; Lee and Yoon, 2018), topic modelling (Budiarto et al., 2021), polarized news detection (Srivastava et al., 2019) and political polarization on Wikipedia (Gode et al., 2023). The model's success on these related tasks suggests that the rich semantic representations of documents that Doc2Vec provides also have the potential to operationalize a metric of politicization.

3. Data and method

USAID's Development Experience Clearinghouse (DEC) represents a rich and largely untapped resource capturing information on aid projects funded by the US government. USAID's evaluation policy (USAID, 2020) stipulates that external evaluations must be carried out for (1) all activities with a total cost exceeding $20 million and (2) each "intermediate result"[2] within a country strategy. The policy further stipulates that plans for the dissemination and use of evaluations must be developed and that evaluation final reports and their summaries must be submitted within three months of completion to the DEC.

Scholars have just begun to leverage the rich information contained in the DEC. For instance, Moore et al. (2023) have developed a standardized taxonomy for benchmarking projects in the agriculture sector. This work lays the foundation for a machine learning algorithm that extracts information on the effectiveness of different interventions and developed standard metrics.

Our study focuses on health projects, for which the DEC contains 4,000 evaluations spanning 70 years. We expect politicization to vary across sectors and activities, arguing that reproductive and maternal healthcare is more politicized than, e.g., malaria control. Following the approach of Moore et al. (2023), we used a balanced sample of 99 reports written from 2003 to 2021 on projects in the health sector.

In selecting the sample, we addressed the limitations inherent in the keyword tagging system of the DEC. Recognizing the frequent inaccuracies

[2]According to USAID's Program Cycle Operational Policy, an intermediate result [IR] is defined as, "A component of a Results Framework in a Mission's CDCS [Country Development Cooperation Strategy]. Intermediate Results are seen as an essential contribution to advancing a DO [Development Objective]. IRs are measurable results that may capture a number of discrete and more specific lower-level results and often define the purpose of projects" (USAID, 2022, p. 127).

in the DEC's keyword-based search functionality, our methodology employed the Development Evidence Large Learning Model (DELLM)[3], a proprietary Large Language Model fine-tuned in collaboration with USAID technical experts. This model demonstrates enhanced capability in accurately categorizing project reports by sector.

The DEC database API was used to operationalize this approach. This integration facilitated an exhaustive analysis wherein DELLM processed the entirety of the DEC's repository to accurately label documents as either 'final evaluations' or 'final grantee reports' within the health sector. Subsequent to this categorization process, a balanced random sampling technique was applied to select a representative subset of 99 labeled reports for further analysis. The sample was balanced to have an even representation of years and countries in which the projects took place.[4]

All but one report are in English and the resulting corpus is 1.6M tokens in size. A vast majority of the reported projects in the sample took place on the African continent but that is representative of the data. Some relevant metadata for the reports is available on the DEC website, most importantly including standardized USAID thesaurus keywords (Donnelly, 2021) for the topics covered in the report. We use these document keywords as labeled data for evaluation.

3.1. Keyword coding

We derived keywords that describe health-related topics from the USAID thesaurus (Donnelly, 2021). The USAID thesaurus keywords are based on 165.000 USAID documents, from across the world, spanning more than 50 years of USAID activities. The USAID thesaurus keywords are commonly used to classify the contents of documents, including USAID project reports (USAID-KSC, 2012). Keywords can be understood as representing the subjects, targets, and interventions of USAID activities. Examples of keywords are 'health', 'HIV/AIDS', and 'bednets'. We derived our keywords from the thesaurus categories relevant to the health sector. Specifically, our keywords are taken from the section 'health and safety' and the 'family planning' sub-section within the 'population and demography' section.[5]

We classify our keywords as politicized (scored 3), non-politicized (1), or potentially politicized (2). Following Carlitz's (2023) notion that the reproductive health sector is more politicized than other sectors, we classified such keywords as politicized. We classify keywords that are not related to reproductive health as non-politicized. Lastly, we classify keywords that capture interventions/targets that can be related to either reproductive or non-reproductive health as potentially politicized. Examples of keywords within the three categories are 'condoms', 'eye diseases', and 'health education'. The classification was done by co-authors with expertise in political science.

3.2. Model

We use Doc2Vec (Le and Mikolov, 2014) in its Gensim (Řehůřek and Sojka, 2010) implementation, trained on the aforementioned DEC corpus, to obtain a potential politicization metric. As our dataset is small for training Doc2Vec, we follow Lau and Baldwin's (2016) approach in initializing Doc2Vec with pretrained word embeddings.[6] The pre-trained word-embedding used is the Common Crawl 300-d vectors with 840b tokens. We chose the DMPV training algorithm which can retain order and thus usually generate better results[7].

We use the model to generate 300-dimensional vectors for each report in the DEC corpus. Based on a list of keywords, it can retrieve the most or least similar documents to the keyword's vectors. We create query vectors by averaging the vectors of query words, with the word vectors coming from the trained model. If the keyword contains more than one word, we split it into single words and take the average vector; we also skip words that are not in the vocabulary of the pre-trained word-embedding model. Using these document embeddings and the hand-coded politicized keywords, we can obtain a potential metric of politicization for a target document by calculating the cosine similarity between the average vector of keywords coded as politicized and the target document.

3.3. Evaluation method

Ideally we would evaluate this approach directly by manually assigning each report a gold standard politicization score and computing the correlation with our metric, but the political scientists in our team consider this an infeasible annotation task due to the abstract nature of the concept.

Instead, we use an indirect 'silver standard' approach based on the report metadata available in the DEC. We score the reports based on whether

[3] https://www.developmetrics.com/our-capabilities/

[4] Sample selection and text extraction from the DEC was performed in collaboration with DevelopMetrics, https://www.developmetrics.com/. None of the authors of this study are affiliated with DevelopMetrics.

[5] Sections K and S14 in the USAID thesaurus.

[6] https://github.com/maohbao/gensim

[7] Model hyperparameters: vector_size: 300; min_count: 1; epochs: 50; dm: 1; seed: 240123. Punctuation and stopwords were removed.

the reports are labeled with politicized keywords in the DEC metadata, and call this the silver score. We then test whether our metric correlates with this silver score, hypothesizing that reports with a higher silver score also get a higher similarity score from our Doc2Vec model. On average, every report has 8 keywords in the DEC metadata, which may be coded differently (scored between 1 and 3 where 3 is politicized, cf. section 3.1). We turn this into a silver standard score by computing the average score of all keywords. If a keyword was not scored by our annotators (e.g. it is not related to the health sector) it gets a score of 1. Documents with a larger proportion of keywords that we coded as politicized thus have a higher silver score.

We consider this a valid evaluation because the Doc2Vec model does not have access to this keyword metadata. The USAID thesaurus keywords are not explicitly listed in the report, although if the keyword is a common word like 'disease', it will be mentioned in the running text. Some more abstract keywords such as 'mass media' do not occur in the report text at all. By receiving an average vector of politicized keywords, the model only has access to our politicization coding at the keyword level, not at the document level. Thus the connection to documents is not given and should be inferred.

4. Results

We compute the cosine similarity between the average politicized keyword vector and the document vectors, using this similarity as our metric. We use the Spearman correlation coefficient to estimate the correlation between our metric and the silver score for all documents. The coefficient obtained is $\rho = 0.280$ with a p-value of 0.005, a weak but statistically significant correlation.

Figure 1 shows all documents ranked by their similarity score plotted against their silver score. This figure shows that top ranked documents on average cover topics that are more politicized according to our annotators, but with some clear deviations from the linear trend around the middle ranks. This suggests that there may be a clustering of documents in the center of the vector space that are not clearly differentiated by politicization.

Among 99 reports, a report on the Mozambique Malaria Program (PA00MGHW) has the highest cosine similarity with politicized keywords. PA00MGHW also has a relatively high silver score of 2.0. While we were initially surprised at a report on a malaria project receiving such a high politicization score, we note that the project included as one of its three main objectives, "Expand access and quality of malaria in pregnancy activities in targeted districts." In the metadata, the report also has keywords related to this topic. This

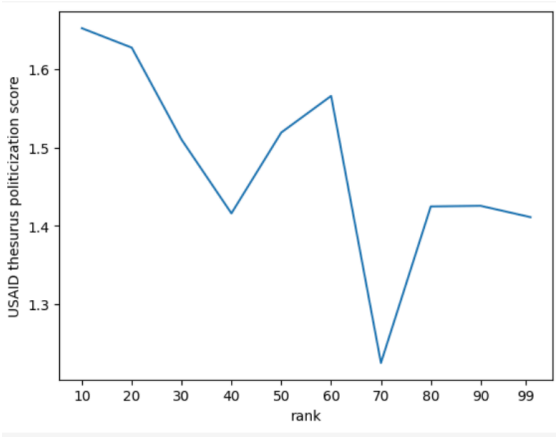


Figure 1: Politicization metric and silver politicization scores, binned in groups of 10 reports.

still lends scope for politicization as we understand it, and points to the importance of going beyond pre-determined keywords. Furthermore, the report also describes a predecessor program more focused on reproductive health, thus influencing the document embedding in a politicized direction, and the report is relatively short. This suggests that segmenting reports into their descriptions of distinct interventions may improve results.

A report on public health training in Ethiopia (PDACG247) has the lowest score on our metric. It also has a silver score of 1, the lowest possible. The main objectives, (1) Development of teaching materials in-country; 2) Strengthen staff through training in pedagogical, supervisory and writing skills; 3) Enhancement of the teaching-learning environment), were indeed not politicized according to our understanding of the concept.

An outlier with the third highest similarity but low silver score is report PA00MGHW. This report appears to be incomplete – that is, the actual evaluation is missing but rather this document is only a series of Annexes, presumably part of a comprehensive evaluation report. Thus, the low silver score indicates what we can miss by relying on externally applied keywords, as the information presented in the annexes does indeed appear to reflect politicized interventions as we understand them (e.g., comprehensive sex education).

A low similarity outlier (rank 97) discusses an Ethiopian reproductive health project, correctly tagged and thus receiving a high silver score of 2.25. The low similarity score was surprising, given the report mentions politicized topics like unsafe abortion. However, the low politicization score may reflect some form of self-censorship and thus may still be capturing a 'real' phenomenon of interest to scholars of politicization.

5. Discussion

While we have shown that our approach yields a metric that correlates with politicized content in foreign aid reports to some extent, there are some clear limitations. First, there is a dependence on manual annotation of politicized keywords. Inducing such keywords from political data sources external to the foreign aid reports would enable easier generalization beyond the health sector. Second, having one vector representation for an entire document proved to be too coarse-grained. Segmenting each report into descriptions of interventions, as also done by Moore et al. (2023), would reduce noise and better represent projects that address a variety of themes. However, as the reports are not consistently structured, this would require manual work. A further limitation is that we were not able to intrinsically evaluate the Doc2Vec model for this domain or perform hyperparameter tuning, due to limited availability of domain-specific resources.

A challenge we encountered throughout our work was coming up with a straightforward conceptualization of politicization that can be grounded in textual data, and identifying documents other than the corpus of reports that we could use to capture politicization. For instance, we searched for policy documents corresponding to Republican vs. Democrat health priorities but failed to find sufficient information. The method is likely more applicable to documents that are more clearly the output of political processes – e.g., comparing political party manifestos to policy documents produced by different parties. In future work we hope to integrate more explicitly political variables to engage more directly with Political Science questions.

The use of static embeddings precludes the possibility of observing different degrees of politicization for the same topics used in different contexts. In much political science work, operationalizations of politicization are conditional on the aid donor and therefore this contextual aspect should be represented in metrics of politicization. Therefore, we propose contextualized embedding-based methods as a future approach. By comparing keywords vector distance in different polarized contexts, we could attribute them a contextual politicization score and develop a politicization metric for keywords at the document level. This metric could be used to study the relation between politicization and project effectiveness. Through its grounding in contextual lexical semantics, this approach could yield deeper insight into the semantic nuances of language used in political discourse and reveal the extent to which political ideologies shape international aid strategies across different donor governments.

Acknowledgements

We are grateful to Lindsey Moore of DevelopMetrics for her help in extracting data from the DEC API. This project was supported by a University of Amsterdam Social and Behavioural Data Science Centre (SoBe DSC) Valorization Grant.

6. Bibliographical References

Asmaa M. Aubaid and Alok Mishra. 2020. A rule-based approach to embedding techniques for text document classification. *Applied Sciences*, 10(11).

Sarah A. Binder. 1999. The dynamics of legislative gridlock, 1947–96. *The American Political Science Review*, 93(3):519–533.

Matteo Bobba and Andrew Powell. 2007. Aid effectiveness: politics matters. Inter-American Development Bank Working Paper #601.

Adam Bonica. 2013. Ideology and interests in the political marketplace. *American Journal of Political Science*, 57(2):294–311.

Adam Bonica. 2014. Mapping the ideological marketplace. *American Journal of Political Science*, 58(2):367–386.

Arif Budiarto, Reza Rahutomo, Hendra Novyantara Putra, Tjeng Wawan Cenggoro, Muhamad Fitra Kacamarga, and Bens Pardamean. 2021. Unsupervised news topic modelling with Doc2Vec and spherical clustering. *Procedia Computer Science*, 179:40–46. 5th International Conference on Computer Science and Computational Intelligence 2020.

Ruth Carlitz. 2023. Aid politics at home and abroad. Downstream consequences for project effectiveness. In *13th Annual Conference of the European Political Science Association*.

Qufei Chen and Marina Sokolova. 2021. Specialists, scientists, and sentiments: Word2Vec and Doc2Vec in analysis of scientific and medical texts. *SN Comput. Sci.*, 2(5).

Sedona Chinn, P. Sol Hart, and Stuart Soroka. 2020. Politicization and polarization in climate change news content, 1985-2017.

Dorottya Demszky, Nikhil Garg, Rob Voigt, J. Zou, M. Gentzkow, James Shapiro, and booktitle = Proceedings of the 17th Annual Conference

of the North American Chapter of the Association for Computational Linguistics (NAACL) Jurafsky, Dan. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings.

Hasibe Dogru, Sahra Tilki, Akhtar Jamil, and Alaa Hameed. 2021. Deep learning-based classification of news texts using doc2vec model. pages 91–96.

Joseph Donnelly. 2021. USAID Thesaurus 2021 dataset. [dataset].

Axel Dreher, Vera Z Eichenauer, and Kai Gehring. 2018. Geopolitics, aid, and growth: The impact of UN Security Council membership on the effectiveness of aid. *The World Bank Economic Review*, 32(2):268–286.

Axel Dreher, Peter Nunnenkamp, and Maya Schmaljohann. 2015. The allocation of German aid: Self-interest and government ideology. *Economics & Politics*, 27(1):160–184.

Samiran Gode, Supreeth Bare, Bhiksha Raj, and Hyungon Yoo. 2023. Understanding political polarization using language models: A dataset and method. *AI Magazine*, 44(3):248–254.

Christopher Kilby and Axel Dreher. 2010. The impact of aid on growth revisited: Do donor motives matter? *Economics Letters*, 107(3):338–340.

Jey Han Lau and Timothy Baldwin. 2016. An empirical evaluation of Doc2Vec with practical insights into document embedding generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 78–86. Association for Computational Linguistics.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Hana Lee and Young Yoon. 2018. Engineering Doc2Vec for automatic classification of product descriptions on O2O applications. *Electronic Commerce Research*, 18:1–24.

Yinghong Liang, Haitao Liu, and Su Zhang. 2020. Micro-blog sentiment classification using Doc2vec + SVM model with data purification. *The Journal of Engineering*, 2020(13):407–410.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Lindsey Moore, Mindel van de Laar, Pui Hang Wong, and Cathal O'Donoghue. 2023. Making impact with agricultural development projects: The use of innovative machine learning methodology to understand the development aid field. *UNU-MERIT Working Paper*, (2023-011).

John J. Nay. 2016. Gov2Vec: Learning distributed representations of institutions and their legal text. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 49–54, Austin, Texas. Association for Computational Linguistics.

Renáta Németh. 2023. A scoping review on the use of natural language processing in research on political polarization: trends and research prospects. *Journal of computational social science*, 6(1):289–313.

Andrew Peterson and Arthur Spirling. 2018. Classification accuracy as a substantive quantity of interest: Measuring polarization in Westminster systems. *Political Analysis*, 26(1):120–128.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Ludovic Rheault and Christopher Cochrane. 2019. Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis*, 28:1–22.

Qianjun Shuai, Yamei Huang, Libiao Jin, and Long Pang. 2018. Sentiment analysis on Chinese hotel reviews with Doc2Vec and classifiers. In *2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pages 1171–1174.

Vertika Srivastava, Ankita Gupta, Divya Prakash, Sudeep Kumar Sahoo, RR Rohit, and Yeon Hyang Kim. 2019. Vernon-fenwick at SemEval-2019 Task 4: Hyperpartisan news detection using lexical and semantic features. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1078–1082.

USAID. 2020. USAID evaluation policy.

USAID. 2022. Program cycle operational policy. ADS Chapter 201.

USAID-KSC. 2012. The USAID thesaurus 6th edition. ID: PN-AEA-100.

Bei Yu, Stefan Kaufmann, and Daniel Diermeier. 2008. Classifying party affiliation from political speech. *Journal of Information Technology & Politics*, 5(1):33–48.

Echo-chambers and Idea Labs: Communication Styles on Twitter

Aleksandra Sorokovikova, Michael Becker, Ivan P. Yamshchikov
Constructor University, Bremen
Technical University of Applied Sciences Würzburg-Schweinfurt (THWS), Würzburg
CAIRO, Technical University of Applied Sciences Würzburg-Schweinfurt (THWS), Würzburg
ivan.yamshchikov@thws.de

Abstract

This paper investigates the communication styles and structures of Twitter (X) communities within the vaccination context. While mainstream research primarily focuses on the echo-chamber phenomenon, wherein certain ideas are reinforced and participants are isolated from opposing opinions, this study reveals the presence of diverse communication styles across various communities. In addition to the communities exhibiting echo-chamber behavior, this research uncovers communities with distinct communication patterns. By shedding light on the nuanced nature of communication within social networks, this study emphasizes the significance of understanding the diversity of perspectives within online communities.

Keywords: echo chambers, computational social science, idea labs

1. Introduction

Online social environments are often characterised by the phenomenon of the so-called echo chambers where participants isolate themselves from opposing opinions, reinforcing their own beliefs through limited communication within their community. These chambers can lead to ad hominem attacks, targeting individuals rather than engaging with their arguments, and straw man arguments, which distort opposing viewpoints for easier dismissal. These dynamics contribute to polarization and the adoption of more extreme positions (Petit et al., 2020). Polarization and echo chambers are commonly observed in social networks, facilitated by recommendation algorithms (Shore et al., 2018; Rossi et al., 2018).

To explore these communities, various algorithms such as the Infomap (Rosvall and Bergstrom, 2008), Louvain algorithm (Campigotto et al., 2014), Stochastic Blockmodels (e.g. (Peixoto, 2020)) or force-directed layout (Gaisbauer et al., 2023) are employed to identify clusters within the user interaction graph. In such setting, users are represented as vertices, and interactions such as social network connections, retweets, and replies on platforms like Twitter (X) are captured as graph edges. These graphs are then partitioned into dense clusters, interpreted as communities sharing similar opinions or engaging in similar activities. Understanding the structure of these communities involves analyzing not only user interactions but also the content they generate. Content analysis allows us to capture the characteristics of the produced content itself (Garimella et al., 2016). Previous studies have investigated echo chambers and polarization in social media, particularly concerning topics like COVID-19, proposing models such as Retweet-BERT and DICE for sentiment analysis and detecting ad hominem attacks (Jiang et al., 2021; Naseem et al., 2020; Delobelle et al., 2019).

In this paper, we focus on identifying Twitter (X) communities related to vaccination. We employ community detection algorithms to identify clusters based on user interactions and the content of their tweets. Additionally, we train classifiers for content analysis, such as sentiment, subjectivity, ad hominem, and straw man arguments. Using these classifiers, we evaluate communication style that characterises each community. This approach enables us to uncover that the patterns of user interaction within communities are clearly different. Moreover, to some extent one could identify community of the user based on their communication style. Thus, we suggest that it is sub-optimal to lump every community under a broad umbrella term "echo-chamber". Instead, we suggest there is a need for a more detailed taxonomy based on the detected systematic differences in the communication styles.

2. Data

For this work a ready-made twitter (X) vaccination dataset[1] was taken, it contains approximately 1.5m tweets on the vaccination topic and 770k unique users. This dataset was collected with TWINT - open-source scraping tool. This dataset is suitable for research for several reasons. Firstly, all tweets relate to the topic of vaccination, and the tweets are taken from a very wide time range (starting in

[1] https://www.kaggle.com/datasets/keplaxo/twitter-vaccination-dataset

Proceedings of the Second Workshop on Natural Language Processing for Political Sciences (PoliticalNLP 2024), pages 91–95

2006, ending in 2019). Secondly, it is possible to build a reply graph, since for tweets that are reply, there is the id of the user to whose tweet this reply was made. A large number of tweets collected in the dataset allows to cover the topic of vaccination from different sides and opinions. Finally, all the discussions are thematically alinged so the differences between texts written by the members of different communities are less prone to be topically aligned. All the discussions are centered around one general topics and the stylistic differences between the texts are more differentiating than the topics these texts discuss. This makes the dataset a good case-study for the core hypothesis of stylistic distinctions between communications styles of various communities.

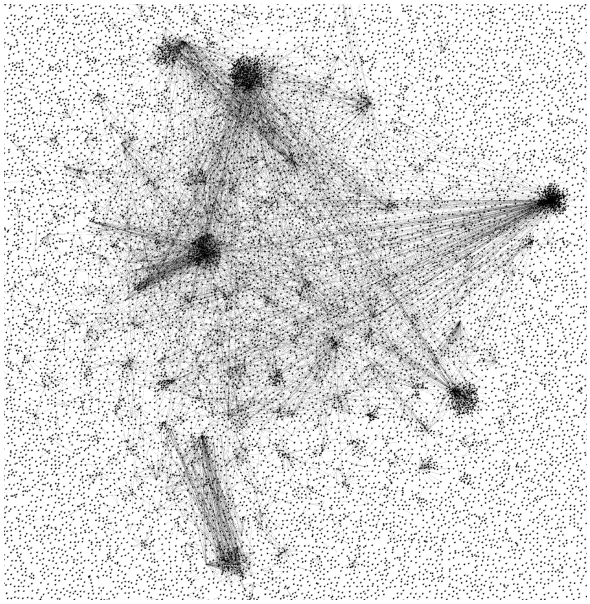


Figure 1: Conversation graph after applying OpenOrd algorithm

3. Experiments

We calculate a variety of metrics using texts of the tweets without any information on the user community in which a given tweet is published. Then we cluster the tweets based on those classifiers and compare mutual information between the obtained text-based clustering and the clusters that are formed in the communication graph. It turns out that the textual information alone partly allows to predict the community in which the user communicates.

3.1. Community clustering

Based on the dataset, a graph was built as follows: each vertex represents a user, and an edge is created if user A made a reply to user B. To get rid of noise and make the graph more dense, a weight was added to the edges, which meant the number of replays of the user. Only edges with weight 3 or higher remained in the final graph. Subsequently, the OpenOrd algorithm (Martin et al., 2011) was used for spatialization, arranging the vertices in two-dimensional space with x and y coordinates, as depicted in Figure 1. Notably, for community detection, the Louvain algorithm was utilized, with clusters determined based on modularity. The computed modularity value for this graph was 0.902, and the subsequent analysis focused on six of the densest communities, informed by clustering coefficient and modularity considerations.

3.2. Text Classification Metrics

We calculate a variety of metrics that assess communication style and are based solely on the textual content of the tweets.

3.2.1. Polarity Scores

The polarity scores were calculated using the Optimized BERT Pretraining Approach (RoBERTa), which was trained on around 58 million tweets. We used the pretrained RoBERTa-based classifier developed in (Barbieri et al., 2020) to calculate the negativity, neutrality and positivity of a give textual input.

3.2.2. Subjectivity

The subjectivity score was determined using the TextBlob (Loria et al., 2018) library. The subjectivity value in TextBlob indicates the degree of subjective or objective content of a given text. Subjectivity refers to how opinionated or subjective the text is, while objectivity refers to a more factual or objective writing style. The subjectivity value is a floating point value between 0.0 and 1.0, where 0.0 indicates a very objective or factual text and 1.0 indicates a very subjective or opinionated text.

3.2.3. Logical fallacy

The two values "label" and "score" were determined using a pre-trained model for logical fallacy detection (Jin et al., 2022). The score label can take values between 0 and 12 and ecodes various logical fallacies, namely: Ad Hominem, Ad Populum, False Causality, Circular Claim, Appeal to Emotion, Fallacy of Relevance, Deductive Fallacy, Intentional Fallacy, Fallacy of Extension, False Dilemma, Fallacy of Credibility, Equivocation. The model assignes a probability between 0 and 1 for every given label.

text/user based cluster	1	2	3	4	5	6
1	445**	942*	172	128	70	58
2	720**	2763*	413	323	239	156
3	262	974*	355**	71	111	71
4	237	827*	94	264**	86	64
5	202	736*	162	82	308**	41
6	205	718*	145	92	71	492**

Table 1: Confusion matrix for text-based clustering used for user-based cluster prediction. The clusters are unbalanced. Cluster number 2 has the biggest support, thus * markes the highers value in a given row, while ** marks the second highest

text/user based cluster	1	3	4	5	6
1	**445**	172	128	70	58
3	262	**355**	71	111	71
4	237	94	**264**	86	64
5	202	162	82	**308**	41
6	205	145	92	71	**492**

Table 2: Confusion matrix for text-based clustering used for interaction-based cluster prediction. Cluster number 2 is removed. The highest values in a given row are marked bold.

4. Predicting Community Membership with Communication Style

Now every tweet could be described by a set of various classifier scores. At the same time we know the community to which the author of the tweet belongs since we have the structure of the reply clusters that we obtained in Section 3.1. To test whether every echo-chamber is characterised with similar communication style we can build clusters of tweets based solely on the classifier scores. Normally, a choice for the number of clusters in a clustering could be difficult. However, since the reply graph clustering procedure has already detected six clusters we can choose six clusters for our text-based clustering as well. Now every tweet belongs to one text-based cluster as well as one reply graph based cluster. Comparing those labels allows us to see to which extent solely textual content of the tweet informs us on the community in which given communication occurs.

Figure 2 demonstrates average scores for six clusters detected in Section 3.1. Once immediately sees that all six are characterize by rather different communication styles. One can see that cluster number two is represented by the point in the "center". It is the biggest cluster that includes a lot of weakly connected users that do not form a dense clique. This cluster has the biggest support in terms of absolute numbers but represents users who are not active members of any on of the five dense communities but are rather occasional posters. Thus, it stands to reason that the average scores of the classifiers for the tweets in this cluster end up in the center of the cloud of points. Table 1 shows the confusion matrix between knn-clusters based on texts of the tweets and the clusters obtained from the graph of interactions between users. The accuracy of the user-based clustering when predicted by text is 0.35 which is quite interesting in itself.

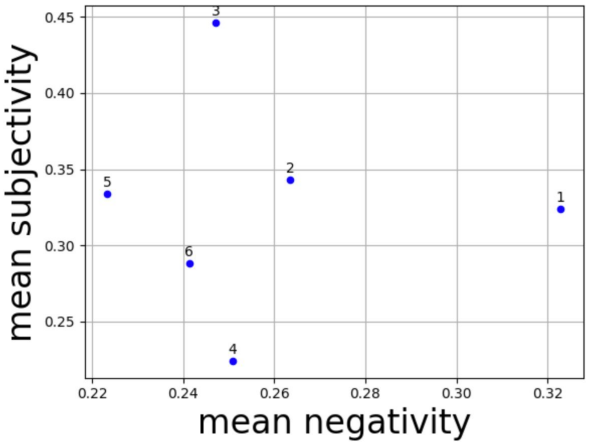


Figure 2: Average scores for mean subjectivity and mean negativity in all 6 user communities. One can clearly see that one of the community is characterised as highly subjective while another is highly negative while scoring lower on subjectivity.

If we drop this cluster number two that represents weakly connected users that do not form a tight community the accuracy reaches 0.44, see Table 2. This highlights that tight communities have have a distinct expressive communication style indeed.

5. Discussion

This study offers an case study and an initial exploration of the varied communication styles within Twitter (X) communities discussing vaccination, challenging the simplistic notion of echo chambers. Through the application of community detection al-

gorithms and text classification metrics, we demonstrate significant diversity in subjectivity, negativity, and logical fallacies across different groups. This suggests a broader spectrum of communicative behaviors in online discussions than typically discussed.

Our findings highlight the importance of nuanced understanding of online discourse, pointing towards the necessity for a more detailed taxonomy of communication styles. This work suggests a pathway for future research to explore the complexities of digital community communication, advocating for a deeper examination beyond conventional categorizations to better understand how these communities interact and evolve.

6. Conclusion

In our investigation of communication patterns within Twitter's (X's) echo chambers, we aimed to identify variations in discourse, specifically seeking environments akin to 'Idea Labs' where open, critical discussion prevails over personal attacks. Utilizing computational methods, we analyzed a substantial dataset to discern these communication styles.

Our results did not confirm the presence of 'Idea Labs' in the studied dataset. However, the study revealed significant variations in communication styles across different echo chambers. Despite discussing identical topics, the textual characteristics within each community were distinct enough to allow for a predictive model to accurately categorize tweets based on their origin.

This finding is critical, demonstrating the extent to which echo chambers can influence discourse style. It also highlights the potential for computational approaches to identify and categorize such patterns in online communication.

Limitations

This paper focuses on one particular case-study. The topic of the discussions is specific and all the results are limited to twitter (X) discussions only. Thus one can not be sure that the results are general and could be applicable to other social media communities or to the discussions around other topics. However, the provided case-study is a good starting point to initiate a deeper discussion of a more nuances approach to community formation in social media that regards the phenomenon in a more holistic manner and takes into account both the structure of the social graph as well as the content of the communications.

Acknowledgements

The authors are deeply thankful to Dr. Felix Glaisbauer for his support, fruitful discussions and valuable advice.

A. Bibliographical References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L_1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.

Romain Campigotto, Patricia Conde-Céspedes, and Jean-Loup Guillaume. 2014. A generalized and adaptive method for community detection.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

James W. Cooley and John W. Tukey. 1965. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301.

Pieter Delobelle, Murilo Cunha, Eric Cano, Jeroen Peperkamp, and Bettina Berendt. 2019. Computational ad hominem detection.

Felix Gaisbauer, Armin Pournaki, Sven Banisch, and Eckehard Olbrich. 2023. Grounding force-directed network layouts with latent space models. *Journal of Computational Social Science*, pages 1–33.

Kiran Garimella, Gianmarco Morales, Aristides Gionis, and Michael Mathioudakis. 2016. Quantifying controversy in social media.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software.

Julie Jiang, Xiang Ren, and Emilio Ferrara. 2021. Social media polarization and echo chambers: A case study of covid-19.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. 2022. Logical fallacy detection.

Steven Loria et al. 2018. textblob documentation. *Release 0.15*, 2(8):269.

Shawn Martin, W. Brown, Richard Klavans, and Kevin Boyack. 2011. Openord: An open-source toolbox for large graph layout.

Usman Naseem, Imran Razzak, Katarzyna Musial, and Muhammad Imran. 2020. Transformer based deep intelligent contextual embedding for twitter sentiment analysis.

Tiago P Peixoto. 2020. Latent poisson models for networks with heterogeneous density. *Physical Review E*, 102(1):012309.

John Petit, Cong Li, and Khudejah Ali. 2020. Fewer people, more flames: How pre-existing beliefs and volume of negative comments impact online news readers' verbal aggression.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Wilbert Samuel Rossi, Jan Polderman, and Paolo Frasca. 2018. The closed loop between opinion formation and personalised recommendations.

Martin Rosvall and Carl T Bergstrom. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the national academy of sciences*, 105(4):1118–1123.

Jesse Shore, Jiye Baek, and Chrysanthos Dellarocas. 2018. Network structure and patterns of information diversity on twitter.

Author Index