A First Step towards Measuring Interdisciplinary Engagement in Scientific Publications: A Case Study on NLP + CSS Research

Alexandria Leto[1] **Shamik Roy**[2*] **Alexander Hoyle**[3]
Daniel E. Acuna[1] **Maria Leonor Pacheco**[1]
[1]University of Colorado Boulder [2]AWS AI Labs [3]University of Maryland
[1]{alexandria.leto, daniel.acuna, maria.pacheco}@colorado.edu
[2]royshami@amazon.com [3]hoyle@umd.edu

Abstract

With the rise in the prevalence of cross-disciplinary research, there is a need to develop methods to characterize its practices. Current computational methods to evaluate interdisciplinary engagement—such as affiliation diversity, keywords, and citation patterns—are insufficient to model the degree of engagement between disciplines, as well as the way in which the complementary expertise of co-authors is harnessed. In this paper, we propose an automated framework to address some of these issues on a large scale. Our framework tracks interdisciplinary citations in scientific articles and models: 1) the section and position in which they appear, and 2) the argumentative role that they play in the writing. To showcase our framework, we perform a preliminary analysis of interdisciplinary engagement in published work at the intersection of natural language processing and computational social science in the last decade.

1 Introduction

As scholarly disciplines have become increasingly specialized, there is a growing need to share knowledge across traditional interdisciplinary boundaries in order to address broader challenges (Vosskamp et al., 1986). Recognizing this trend, scholarly institutions have established interdisciplinary centers (Turner et al., 2015; Leahey and Barringer, 2020) and funding agencies are incentivizing such collaborative efforts (Hackett, 2000; Rhoten and Parker, 2024). However, it is unclear how this interdisciplinarity is reflected in the main outcome of science: scientific publications.

Existing measures of interdisciplinarity are overly simplistic and fail to capture the depth of cross-disciplinary integration in research (McCarthy and Dore, 2023). Current metrics, such as affiliation diversity, keywords, and citation patterns (Porter and Rafols, 2009; Van Noorden, 2015; Chen et al., 2021) often overlook how co-authors' expertise is used, and keywords fail to truly reflect a paper's content. Moreover, where and how a citation is made within papers remains largely unexplored. For example, citing papers from other fields in the opening paragraph of the introduction might signal higher interdisciplinarity than doing so in a footnote of the method section. In turn, using these references to ground findings and substantiate claims signals a deeper cross-disciplinary engagement than briefly discussing the connections between two bodies of work.

To address this challenge, we propose an automated framework for characterizing and measuring the degree of meaningful cross-disciplinary engagement in scientific publications by explicitly modeling the structure (i.e., where do interdisciplinary references appear?) and content (i.e., how are these references used to make claims?) of scientific papers. To do this, we track all interdisciplinary citations in a given article. We consider a citation to be interdisciplinary if it belongs to a venue outside of the current paper discipline. Then, for each sentence containing an interdisciplinary citation, we identify its argumentation type. To define argumentation types we build on the schema proposed by Lauscher et al. (2018), which distinguishes between claims made about the author's own work, claims made about the background of the author's work, and factual statements that serve as evidence for or against a claim. Details about the implementation of this framework are provided in Sec. 3.

As a case study, we apply our framework to research papers in the interdisciplinary field of natural language processing (NLP) and computational social science (CSS). In the past decade, greater volumes of text data and more accessible methods have caused a growth in popularity of this research area (Grimmer and Stewart, 2013), including a

*Work done before joining AWS AI Labs.

Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 144–158

dedicated workshop in leading NLP conferences*. However, like other interdisciplinary research, the synergy between NLP and CSS is often violated due to a misalignment between the research outlook, requirements, and expertise of the researchers in the corresponding fields. In recent work, McCarthy and Dore (2023) manually review a set of CSS papers that incorporate text analysis methods, and conclude that many of these contributions present what they call *descriptive findings*: papers that present descriptive catalogs of evidence derived from analyzing social data (e.g., word distributions found in tweets about mass shootings) but that fail to integrate these findings with relevant social science theory. They contrast these contributions with *integrative findings*, which seek to achieve synergistic methodology to meet the standards of both disciplines, furthering theory. Similarly, Baden et al. (2022) note that available NLP methods often fail to meet the needs of social science research, where a limited ability to incorporate theory damages methods' validity. Our work presents the first large-scale analysis of the way in which authors working on NLP+CSS have engaged with the literature in fields outside of computer science and linguistics over the past 10 years.

We make the following contributions: 1) We propose a first step towards a general computational framework to analyze interdisciplinary engagement in scientific publications. 2) We construct a comprehensive dataset of computational social science articles published in NLP venues in the last 10 years. 3) We perform a large scale analysis of interdisciplinary engagement in NLP+CSS research, and show that while interest in NLP+CSS work is growing, there is a decreasing trend in the engagement with outside disciplines in the mainstream NLP conferences. We explore these trends in the context of the main topics of interest in the NLP+CSS community and how they have shifted over time. We also show that dedicated workshops like the *NLP and Computational Social Science Workshop* attract highly interdisciplinary contributions, fulfilling their mission of providing an outlet for this type of work.

2 Related Work

Most previous studies at the intersection of NLP and the Science of Science have analyzed scientific publications by looking at their citation patterns. Some of this work has focused on the way citation behavior relates to the scientific content of articles. For example, Jurgens et al. (2018) studied the effect of framing contributions through citations, Qazvinian and Radev (2008) incorporated citation networks in document summarization, and Cohan et al. (2020) used citation graphs to learn scientific document embeddings. Another line of research has studied citation behavior in the NLP literature, by looking at how scientific articles are distributed across geographies (Rungta et al., 2022), or over different types of NLP papers (short, long, demo, etc.) (Mohammad, 2020).

While language-centered approaches are scarce in science of science research, there have been some prior efforts in this direction. Some notable examples are: studying cross-field jargon interpretation (Lucy et al., 2023), the influence of articles in the scientific community (Yogatama et al., 2011; McKeown et al., 2016; Gerow et al., 2018), the evolution of scientific topics (Prabhakaran et al., 2016), and the prevalence of different research themes (Mendoza et al., 2022).

In this paper, we look at *when* and *how* interdisciplinary citations are used in scientific articles. Previous work looking at citation context in scientific discourse has modeled the sentiment towards cited articles (Athar and Teufel, 2012; Munkhdalai et al., 2016), citation intent (Kunnath et al., 2022), purpose and influence (N. Kunnath et al., 2021), and critical vs non-critical arguments (Te et al., 2022). In our study, we focus on the location and argumentation role of interdisciplinary citations.

Our work is broadly related to the argumentation mining literature (Peldszus and Stede, 2013; Lawrence and Reed, 2020). We study argumentation in the context of scientific publications. While previous studies focus on identifying argumentative discourse units (Binder et al., 2022) and their relations (Lauscher et al., 2018; Gao et al., 2022), we study how work coming from outside disciplines is used to make arguments in scientific articles.

3 Framework

In this section, we describe our automated framework to model the content and structure of NLP+CSS papers. Our framework is composed of three sub-tasks: 1) Identifying papers that present CSS findings and contributions, 2) For each relevant paper, identifying all cross-disciplinary citations, and 3) For each cross-disciplinary citation,

*https://aclanthology.org/venues/nlpcss/

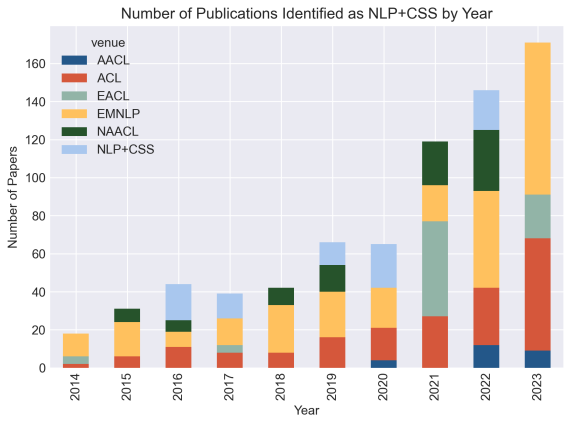

Figure 1: Resulting dataset of NLP+CSS publications over the last decade

identifying the type of argument being made.

3.1 Data Collection

To perform our analysis, we first need to construct a dataset of representative CSS articles published in NLP conferences. To do this, we first collect all long papers published in the ACL Anthology between 2014 and 2024 in all main NLP conference proceedings (ACL, EMNLP, NAACL, EACL and AACL), as well the Workshop on NLP and Computational Social Science (NLP+CSS). This results in a dataset of about 13,000 scientific papers.

Next, we need to identify which of those 13,000 papers correspond to CSS contributions. In the next section, we outline our method to achieve this.

3.2 CSS Track Identification

With the goal of building a classifier to identify CSS papers, we label a subset of about 1,800 example papers with their track to serve as training data. Namely, we collect all papers that were listed under the "Computational Social Science and Cultural Analytics" track in schedules and conference handbooks available online for the years of 2021, 2022 and 2023 and consider them as positive examples. We also add all papers published in the NLP+CSS Workshop proceedings in any given year to the set of positive examples. To generate negative examples, we follow the same procedure for each additional track (e.g., Language Generation, Machine Translation). This results in 236 positive and 1,552 negative examples.

We build a binary classifier to automatically identify papers within the dataset that fall under NLP+CSS, and are thus pertinent to our analysis. To do so, we fine-tune a pre-trained RoBERTa classifier using the abstracts of the labeled subset of data. General statistics for the resulting dataset are shown in Fig. 1, and a more detailed dataset breakdown may be found in App. A.1.

3.3 Cross-Disciplinary Citation Mapping

To study interdisciplinary engagement, we need to track all citations in papers that reference work outside of NLP, Linguistics and Computer Science. We are particularly interested in identifying the sections in the paper where these citations appear, as we hypothesize that this might signal different types of interdisciplinary engagement.

To do this, we extract the content of the each article using Grobid [†]. Then, we use the SciPDF Parser [‡] to convert the extracted content into a dictionary format including the article text and citations, as well as section breakdowns. We model each publication as a set of canonical section headers typical of NLP papers paired with their corresponding in-text citations. We consider the following canonical sections: "Introduction", "Related Work", "Method", "Experiments", "Conclusion" and "Appendix". To arrive at this canonical section breakdown, the sections initially parsed by Grobid are assigned to one of these headers using a rule-based approach based on string matching. Then, all in-text citations within a section are mapped to the corresponding "canonical" section. Details about this process are outlined in Appendix A.2.

In addition, information for each entry in the "References" section of a publication (title, journal, publication data, id, etc.) is extracted. For all references with an available, valid id (DOI, arXiv, or url), we query the Semantic Scholar API (Kinney et al., 2023) for a "field of study". To handle entries with no valid id, we match venues to the appropriate field of study using Google Scholar's lists of venues per field. All remaining unassigned venues were annotated manually by the authors of the paper. Finally, all in-text citations were mapped to reference entries by string matching the author name and the publication year.

3.4 Argument Analysis

Once we have identified all interdisciplinary references, we are interested in modeling the argumentative role they play in the article. To do this, we build on the argumentation schema proposed by Lauscher et al. (2018). This schema builds on the

[†] https://github.com/kermitt2/grobid
[‡] https://github.com/titipata/sci_pdf_parser

Toulmin model (Toulmin, 2003), one of the most widely used theoretical frameworks of argumentation. The Toulmin model was originally conceived for the legal domain, and recognizes six types of argumentative components: claim, data, warrant, backing, qualifier, and rebuttal. Lauscher et al. (2018) do an annotation study on scientific documents, and simplify the Toulmin model by focusing only on claims and data. Further, they break down claims into *own claims* and *background claims* to differentiate between claims that relate to the author's own work, and claims that relate to the background of the author's work (e.g., about related work or common practices).

We build a classifier to identify whether interdisciplinary references correspond to *data* (e.g., fact or example that serve as evidence for or against a claim), *background claims*, or *own claims*. To do so, we fine-tune a pre-trained BERT classifier using the dataset provided by Lauscher et al. (2018). This dataset is comprised of 13,592 arguments: 6,004 own claims 3,291 background claims, and 4,297 data statements. Upon closer inspection of the dataset, we found that many examples correspond to figure titles and standalone citations. To deal with this, we removed all training examples with no verb phrases using spaCy. This results in 8,737 arguments: 4,968 own claims, 2,613 background claims, and 1,156 data statements.

4 Experiments and Analysis

To validate our framework, we first perform an experimental evaluation of each component. Then, we use the full framework to perform an extensive analysis of interdisiciplinary engagement for our full dataset of CSS+NLP papers.

4.1 Framework Evaluation

To train and evaluate the track identification and argument type classifiers, we performed 5-fold cross-validation and trained using the AdamW optimizer, the cross-entropy loss, and a learning rate of $1e-5$. For early stopping, we used the macro F1 on the validation set, consisting of 20% of the training examples for each fold.

We present the resulting metrics of our track classifier in Tab. 1. We obtain strong performance for this task considering the highly imbalanced nature of the data. This suggests that we can trust that our dataset of NLP+CSS papers is representative of the real distribution. Detailed results per fold are shown in Tab. 8 in Appendix A.3.

Class	Precision	Recall	F1
CSS	0.8189	0.8326	0.8257
Not CSS	0.9742	0.9716	0.9729
Macro Avg	0.8965	0.9021	0.8993

Table 1: Avg. Results for Track Prediction

Class	Precision	Recall	F1
Own Claim	0.7460	0.7693	0.7564
Background Claim	0.6289	0.6184	0.6218
Data	0.5233	0.4669	0.4745
Macro Avg	0.6327	0.6182	0.6176

Table 2: Avg. Results for Argument Type Prediction

We present the resulting metrics of our argument type classifier in Tab. 2. We obtain relatively good performance for the two types of claims, but struggle with data statements. This is a challenging task, and our results are in line with the skewed nature of the dataset, where there is considerably less support for data examples. Detailed results per fold are shown in Tab. 10 in Appendix A.3.

4.2 Analysis of Interdisciplinary Engagement

In this section, we use the framework introduced in Sec. 3 to perform an exhaustive analysis of the engagement of NLP+CSS papers with work outside of NLP, Computer Science and Linguistics. To do this, we first used the CSS track classifier to derive the dataset presented in Fig 1. Then, we ran the citation mapping procedure. Next, for every sentence involving or preceding a citation or reference, we predict its argument type using our argument type classifier. To train the final argument type classifier, we used the full dataset of arguments from Lauscher et al. (2018), containing all five folds. Finally, we model 15 topics in the abstracts of the NLP+CSS papers to identify growing and shrinking trends.

The final dataset is comprised of 741 NLP+CSS papers, published across five NLP conferences and one workshop, and spanning 9 years (2014-2023). Within these 741 publications, we have a total of 16,652 references annotated with the canonical section in which they appear, their scientific discipline and their predicted argument type.

Below, we present our analysis organized by the

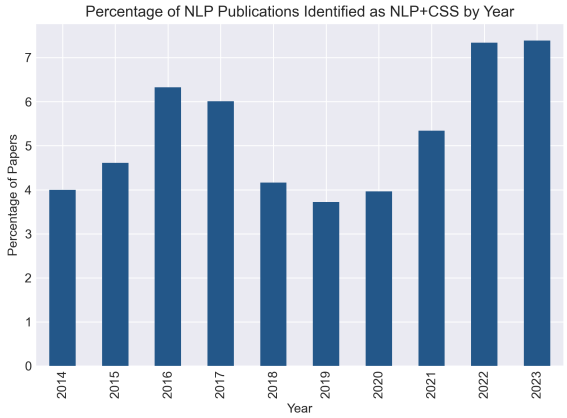

Figure 2: Percentage of all gathered NLP papers that were predicted or labelled as NLP+CSS papers

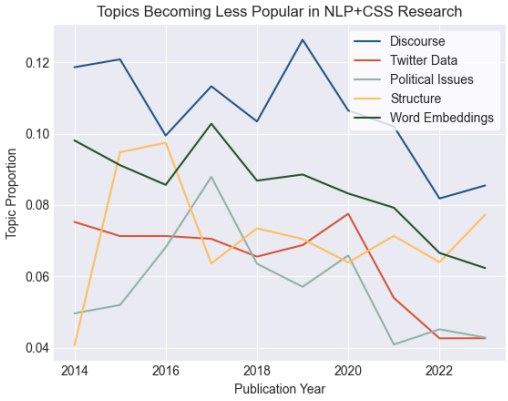

Figure 3: Topics becoming less popular in NLP+CSS research

main trends that we observed in the data.

NLP+CSS research is growing. Fig. 2 shows the number of NLP+CSS papers per year in all five NLP conferences. In this figure, we observe that the percentage of NLP work identifying as NLP+CSS research has significantly increased in the last three years. This growth comes at a time where the volume of NLP work in general has significantly increased. We also observe a peak in NLP+CSS work in 2016 and 2017.

Rising and falling topic trends might explain NLP+CSS prominence. To gain a better understanding of shifts in content of the NLP+CSS papers, we conducted topic modelling with gibbs sampling over all abstracts using tomotopy (Lee, 2022). We modelled 15 topics and eliminated the 50 most-common words. Based on the top 10 words associated with each topic, as well as the top documents associated with each topic, we manually assigned the topic a title. Details for each of the 15 topics, including associated words and assigned title are presented in App. A.5.

We used topic distribution information to understand shifts in topics over the years. We identify five topics that have become less prevalent over the last five years (Fig. 3), and three topics that have become increasingly popular over the full 9 year period (Fig. 4). First, we can appreciate that the 2016-2017 peak in NLP+CSS papers corresponds with increased interest in topics related to politics, public discourse and hate speech. We hypothesize that these trends could be related to the U.S. general election and the uptake in political discourse on social media. A similar peak can be seen for political issues around 2020, when the next U.S. general election occurred.

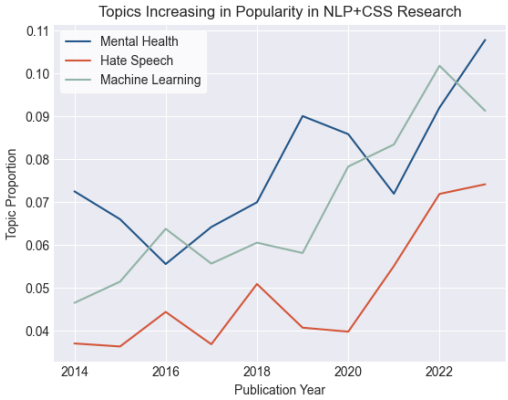

Figure 4: Topics becoming more popular in NLP+CSS research

We also observe a decline in techniques like structured prediction after 2016, when neural networks like LSTMs became more popular. Word embeddings experienced a pronunced peak between 2018 and 2020–coinciding with the rise of contextualized word representations–and it has been decreasing since then. We hypothesize that this is influenced by a shift to LLMs and generative solutions. On the other hand, general machine learning vocabulary has been steadily increasing in the last 9 years. This is unsurprising, as NLP research has become increasingly more entangled with machine learning research. However, this has not caused a shift in focus away from social topics, as we also observe a steady incline in research around mental health and hate speech.

Finally, we observe a sharp decrease in papers dealing with Twitter data in the last two years. This coincides with changes in leadership at Twitter, and

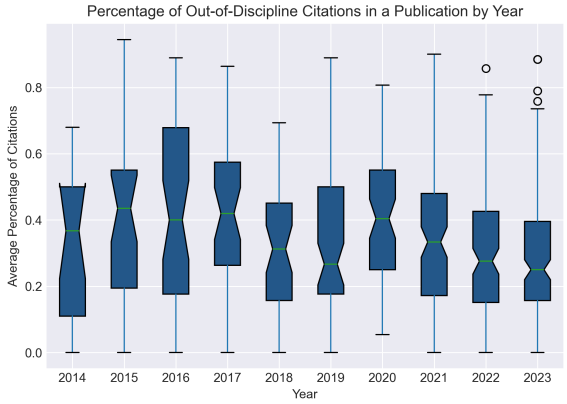

Figure 5: Average % of Out-of-Discipline citations in NLP+CSS papers per year

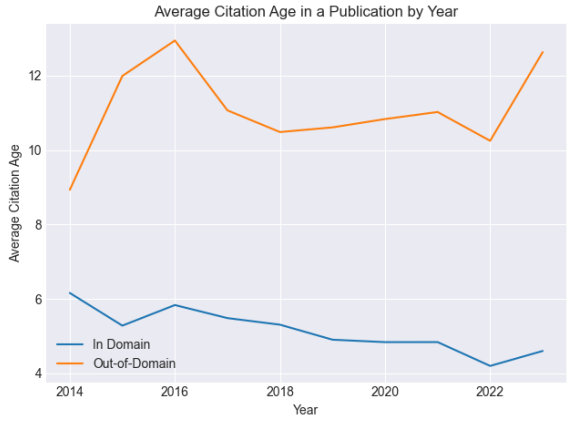

Figure 6: The average age of Out-of-Discipline citations are significantly older at the time of reference

the increasing difficulty in accessing the data.

Cross-disciplinary engagement in NLP+CSS articles has decreased in the last three years. While the number of NLP+CSS papers has grown in the last five years, the proportion of out-of-discipline and specifically social scientific papers cited has decreased in the same time span (Figs 5 and 7). Initially, peaks in the amount of cross-disciplinary engagement coincide with peaks in the prevalence of NLP+CSS work (around 2016 and 2020, as seen on Fig. 2). However, after 2020, the trends seem to be in opposition to each other - more NLP+CSS work but less interdisciplinary engagement. Interestingly, the decrease in proportion of interdisciplinary citations coincides with the LLM boom in NLP research.

To further inspect this trend, we look at the most cited papers in NLP+CSS work for each year (Tab 3). The most cited papers prior to 2018 had to do with social-adjacent topics such as dialog (Prabhakaran and Rambow, 2014) and language in social media data (Mitchell et al., 2015), and were not as frequently cited (cited in 11-15 in-text citations). Starting in 2018, top-cited papers include a survey of affective computing (Poria et al., 2017), a paper about deep neural networks (Alzantot et al., 2018), and pre-trained language models (Liu et al., 2019). The amount of papers citing them significantly grew (20-90 in-text citations).

Out-of-Discipline citations are older. In plotting the average age of Out-of-Discipline versus In-discipline citations at the age of reference (Fig. 6), we find that Out-of-Discipline citations are significantly older. This may communicate a tendency to engage only with more well-known, seminal

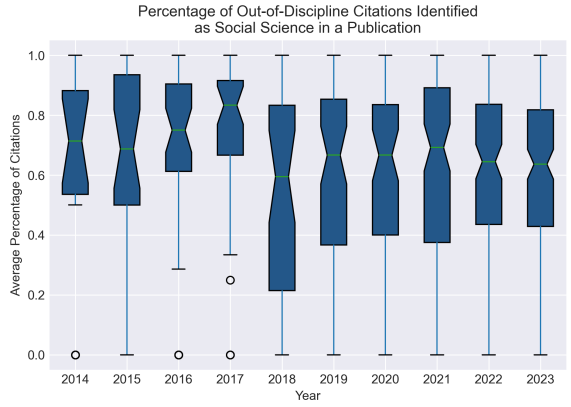

Figure 7: Average percentage of Out-of-Discipline citations within a publication that are considered Social Science-specific

papers. It is also illustrative of the much shorter "life-span" of AI-adjacent research. Over the past decade, this has become increasingly true (Singh et al., 2023; Nguyen and Eger, 2024).

Dedicated workshops are more interdisciplinary. In Fig. 8 we show boxplots for the proportion of interdisciplinary citations per venue. We observe no meaningful difference in the interdisciplinary citation patterns of CSS papers between the different NLP conferences, with the exception of AACL, which appears to be less interdisciplinary. This could be due to the fact that AACL is a new venue, and therefore attracts more traditional NLP work.

However, we find a significant increase in interdisciplinary citations for the NLP+CSS workshop. This suggests that dedicated workshops attract more interdisciplinary work, which is in line with the mission of the workshop series: to foster the progress of CSS, and to integrate CSS with

Year	Most Cited Paper	Times Cited
2014	Predicting Power Relations between Participants in Written Dialog from a Single Thread (Prabhakaran and Rambow, 2014)	11
2015	Exploiting Similarities among Languages for Machine Translation (Mikolov et al., 2013)	10
2016	Inferring Latent User Properties from Texts Published in Social Media (Volkova et al., 2015)	15
2017	Quantifying the Language of Schizophrenia in Social Media (Mitchell et al., 2015)	12
2018	A review of affective computing: From unimodal analysis to multimodal fusion (Poria et al., 2017)	16
2019	Generating Natural Language Adversarial Examples (Alzantot et al., 2018)	19
2020	RoBERTa: A Robustly Optimized BERT Pretraining Approach (Liu et al., 2019)	20
2021	RoBERTa: A Robustly Optimized BERT Pretraining Approach (Liu et al., 2019)	39
2022	RoBERTa: A Robustly Optimized BERT Pretraining Approach (Liu et al., 2019)	61
2023	RoBERTa: A Robustly Optimized BERT Pretraining Approach (Liu et al., 2019)	90

Table 3: Most cited paper per year

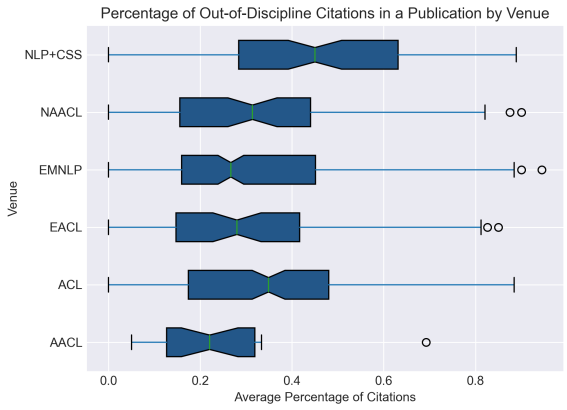

Figure 8: % of Out-of-Discipline citations per venue

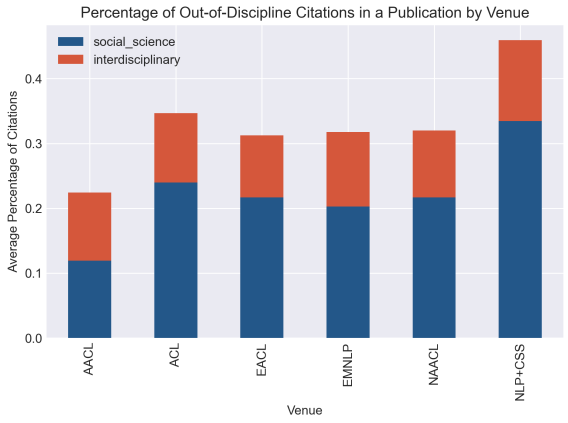

Figure 9: % Out-of-Discipline per venue

current trends and techniques in NLP.

Social science papers are the most commonly cited. Unsurprisingly, among the number of out-of-discipline papers cited in CSS papers, the vast majority correspond to social science venues. This can be observed in Fig. 7 (prop. of social science papers among out-of-discipline papers per year) and Fig. 9 (prop. of social science papers among out-of-discipline papers per venue). Moreover, we find that the trend is very stable among different years and venues.To perform this analysis, we consider the social science and humanities fields outlined in App A.4.

Finally, to further investigate which social science (and humanities) fields are most commonly cited, we plot the top-10 fields in Fig 10. We find that *psychology* is the top most cited social science / humanities field. It is followed by *political science*, *general social science*, *sociology*, *business*, *economics*, *communication*, and *education*.

In-discipline references are used more often for *own claims* and *data statements*. In Tab. 4 we can appreciate the differences in argument type between in-discipline and out-of-discipline references. We find that when making claims that relate to the author's own work (own claims), and stating facts or examples that serve as evidence, authors more often reference work within the same discipline. Conversely, when making claims that relate to the background of their work (background claims), authors more often reference work outside of their discipline. We also note that this difference is even more pronounced in Social Science-specific Out-of-Discipline citations. While this is an unsurprising result, it is interesting that there is still a significant amount of out-of-discipline citations used to make *own claims*, which signals meaningful interdisciplinary engagement.

Citation Type	Background Claim	Data	Own Claim
In-Discipline	53.98%	6.92%	39.1%
Out-of-Discipline	64.16%	4.77%	31.06%
Social Science	68.91%	4.13%	26.96%

Table 4: Percentage of Argument Types Supported by In-Discipline, Out-of-Discipline and Social Science-specific Out-of-Discipline Citations

Sections matter when referencing out-of-discipline work. In Fig. 11 we can appreciate significant differences in argument types and citation

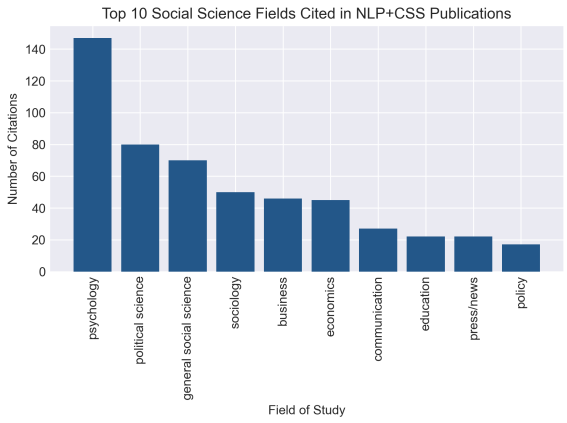

Figure 10: Top 10 most cited social science fields

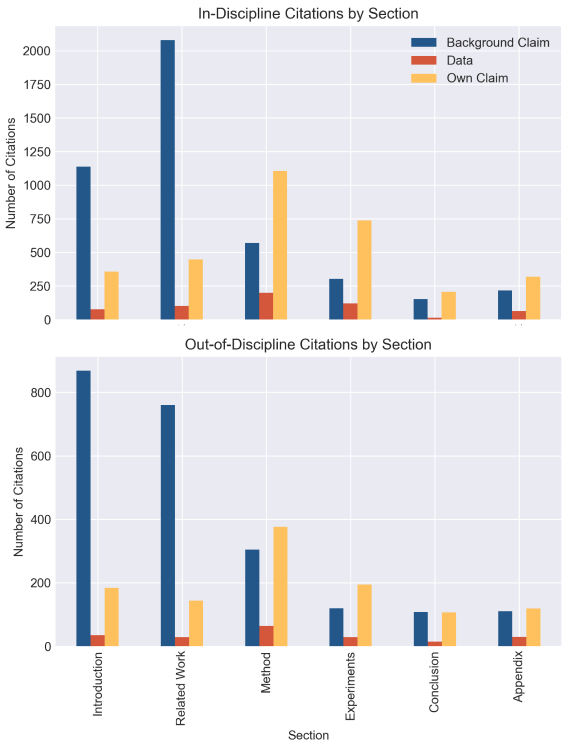

Figure 11: Argument Types Supported by In-discipline vs. Out-of-Discipline citations per Section

frequency by section. The number of background claims made in the introduction increases significantly for out-of-discipline references. This makes sense for CSS papers that are attempting to frame and motivate their contributions with respect to the broader social science literature. Additionally, we can see that for both method and experiments sections, in-discipline references are used more often to back *own claims*, while out-of-discipline references are used more evenly to support both *own claims* and *background claims*.

Papers with higher rates of Out-of-Discipline citations are integrative. We conduct a qualitative analysis of the three papers identified as having the greatest proportion of Out-of-Domain references and compare to the three papers with the lowest proportion. Our observations for each paper are summarized in Tab. 5.

We find that the papers identified with the most Out-of-Discipline citations each seek to build upon existing social science work, tying relevant theory strongly into their motivation, methods, and discussion of results. Following (McCarthy and Dore, 2023), we are inclined to describe them as "integrative" papers. Alternatively, we find that papers with the lowest rates of Out-of-Discipline citations are mainly method papers grounded in computer science research. Each of these papers addresses a relevant social issue or task, but the main focus is on formulating a prediction task, proposing a computational model and analyzing the prediction performance.

5 Discussion and Future Work

We emphasize that this paper represents only a first step toward our envisioned framework. Going forward, we will expand our labelled dataset for greater coverage of additional tracks and workshops typical of NLP conferences. We aim to predict, with high confidence, what track a paper was submitted to, allowing for a more holistic analysis of the field of NLP and its relationship with out-of-discipline work. We will conduct both general and per-track analyses to identify additional trends and explore whether the trends observed in NLP+CSS hold true in other types of work.

We are particularly interested in studying the lasting power of NLP work and how that has shifted under recent advancements in the field. We are also interested in expanding on (McCarthy and Dore, 2023) to further investigate the differences in integrative and descriptive work and their engagement patterns with out-of-discipline scholarship in the field of NLP as a whole. We seek to provide examples of how integrative work may be carried out in the age of LLMs.

In addition to expanding beyond NLP+CSS, we are also interested in improving on our argument analysis component. We want to experiment with alternative frameworks, as well as with modeling argumentative relations between claims. Moreover, we will further investigate our hypothesis that the

Most Interdisciplinary	
(Wu et al., 2023)	This paper presents a study of values, morals and gender bias in a dataset of folk tales spanning cultures. Most experiments are designed to reinforce a relevant theoretical claim.
(Yang et al., 2015)	This is a study of the relationship between personal traits and brand preference. The method is grounded in theory and seeks to investigate a debated topic in marketing research through a large scale study.
(CH-Wang and Jurgens, 2021)	This paper studies differences in word choices for significant others and indefinite people to examine shifts in societal attitudes toward sexuality and gender. Motivation for their method and discussion of results relies heavily on theoretical background.
Least Interdisciplinary	
(Yang et al., 2022)	This paper presents a method for using facial emotions to extract sentiment from sentence-image pairs.
(Zhou et al., 2016)	This paper studies enhancing limited personal usage information with the goal of improved search personalization. They mainly call on previous computational methods to motivate deign choices.
(Li et al., 2023)	This paper examines the effectiveness of identifying stance in social media posts given background knowledge about the topic. Their proposed method is a variant of Retrieval Augmented ChatGPT.

Table 5: Summary of observations from analysis of NLP+CSS publications with greatest and least proportion of Out-of-Discipline citations

location of a citation within a paper may signal higher or lower engagement. Conducting qualitative analysis with a science-of-science expert will allow us to tie these observations to meaningful differences in engagement. We envision combining this information with the argumentation framework to make deeper claims about the way citations are used within a publication. We also recognize the need to integrate and contrast our framework with recent Science of Science techniques such as citation network analysis.

6 Conclusion

We recognize a need to evaluate interdisciplinary research due to its growing popularity paired with a lack of sufficient methods for studying engagement between disciplines. In line with this, we propose a general scalable framework for tracking interdisciplinary citations within publications. Our framework allows for tracking the section where interdisciplinary citations appear and the argumentative role they play within a publication.

To showcase our framework, we performed a large scale analysis on the interdisciplinary engagement of research in the field of natural language processing and computational social science. To do this, we constructed a comprehensive dataset of NLP+CSS papers published in the NLP venues in the last decade. Our analysis revealed several trends, including a growth in the number of NLP+CSS publications, compounded with a decrease in cross-disciplinary engagement in NLP+CSS papers coinciding with the advent of LLMs. These findings are in line with previous work highlighting the gaps between the two fields (Baden et al., 2022; McCarthy and Dore, 2023). However, we also find that dedicated workshops, such as the *NLP and Computational Social Science workshop*, attract contributions that exhibit higher engagement with the social scientific literature.

7 Limitations

The work presented in this paper has four main limitations: (1) We defined interdisciplinary references as those that cited a paper outside of NLP, Computer Science or Linguistics. We recognize that this is a simplification and that scientific contributions can vary widely within certain venues. (2) The classifiers used to identify argument types was trained and evaluated on out-of-domain data. While this data was also comprised of scientific articles, some domain drift is to be expected when moving from the computer graphics domain to the natural language processing domain. A post-hoc manual evaluation is needed to check and establish the performance for our dataset. (3) We complemented our citation analysis with a topic analysis to tie the findings to some of the most prominent research trends in the literature. We recognize the limitation of topic models to accurately capture this type of information accurately. However, we believe that this risk is diminished when looking at aggregated trends, rather than at individual mapping between papers and topics. (4) The fact that we are using automated techniques for the analysis necessarily carries some uncertainty. Even if we were to improve our models considerably, our large-scale analysis has a margin of error. It is important to acknowledge this when presenting our

findings.

8 Ethical Considerations

To the best of our knowledge, no code of ethics was violated during the development of this project. We used publicly available tools and data to develop our framework and perform our analysis. We reported all pre-processing steps, learning configurations, hyperparameters, and additional technical details. Due to space constraints, some information was relegated to the Appendix. The results reported in this paper support our claims and we believe that they are reproducible. The analysis reported in Section 4.2 was done using the outputs of matching algorithms and machine learning techniques and do not represent the authors personal views. The uncertainty of our predictions was adequately acknowledged in the Limitations Section, and the estimated accuracy was reported in Sec. 4.1.

References

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

Awais Athar and Simone Teufel. 2012. Context-enhanced citation sentiment detection. In *Proceedings of the 2012 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*, pages 597–601.

Christian Baden, Christian Pipal, Martijn Schoonvelde, and Mariken A. C. G van der Velden. 2022. Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda. *Communication Methods and Measures*, 16(1):1–18.

Arne Binder, Leonhard Hennig, and Bhuvanesh Verma. 2022. Full-text argumentation mining on scientific publications. In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*, pages 54–66, Online. Association for Computational Linguistics.

Sky CH-Wang and David Jurgens. 2021. Using sociolinguistic variables to reveal changing attitudes towards sexuality and gender. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9918–9938, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shiji Chen, Junping Qiu, Clément Arsenault, and Vincent Larivière. 2021. Exploring the interdisciplinarity patterns of highly cited papers. *Journal of Informetrics*, 15(1):101124.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282.

Yingqiang Gao, Nianlong Gu, Jessica Lam, and Richard H.R. Hahnloser. 2022. Do discourse indicators reflect the main arguments in scientific papers? In *Proceedings of the 9th Workshop on Argument Mining*, pages 34–50, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Aaron Gerow, Yuening Hu, Jordan Boyd-Graber, David M Blei, and James A Evans. 2018. Measuring discursive influence across scholarship. *Proceedings of the national academy of sciences*, 115(13):3308–3313.

Justin Grimmer and Brandon M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297.

Edward J. Hackett. 2000. 12. interdisciplinary research initiatives at the u.s. national science foundation. In Peter Weingart and Nico Stehr, editors, *Practising Interdisciplinarity*, pages 248–259. University of Toronto Press.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

Rodney Michael Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David W. Graham, F.Q. Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler Murray, Christopher Newell, Smita R Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, A. Tanaka, Alex D Wade, Linda M. Wagner, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine van Zuylen, and Daniel S. Weld. 2023. The semantic scholar open data platform. *ArXiv*, abs/2301.10140.

Suchetha Nambanoor Kunnath, David Pride, and Petr Knoth. 2022. Dynamic context extraction for citation classification. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International*

Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 539–549.

Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. 2018. An argument-annotated corpus of scientific publications. In *Proceedings of the 5th Workshop on Argument Mining*, pages 40–46, Brussels, Belgium. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Erin Leahey and Sondra N. Barringer. 2020. Universities' commitment to interdisciplinary research: To what end? *Research Policy*, 49(2):103910.

Minchul Lee. 2022. bab2min/tomotopy: 0.12.3.

Ang Li, Bin Liang, Jingqian Zhao, Bowen Zhang, Min Yang, and Ruifeng Xu. 2023. Stance detection on social media with background knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15703–15717, Singapore. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Li Lucy, Jesse Dodge, David Bamman, and Katherine Keith. 2023. Words as gatekeepers: Measuring discipline-specific terms and meanings in scholarly publications. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6929–6947, Toronto, Canada. Association for Computational Linguistics.

Arya D. McCarthy and Giovanna Maria Dora Dore. 2023. Theory-grounded computational text analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1586–1594, Toronto, Canada. Association for Computational Linguistics.

Kathy McKeown, Hal Daume III, Snigdha Chaturvedi, John Paparrizos, Kapil Thadani, Pablo Barrio, Or Biran, Suvarna Bothe, Michael Collins, Kenneth R Fleischmann, et al. 2016. Predicting the impact of scientific concepts using full-text features. *Journal of the Association for Information Science and Technology*, 67(11):2684–2696.

Óscar E Mendoza, Wojciech Kusa, Alaa El-Ebshihy, Ronin Wu, David Pride, Petr Knoth, Drahomira Herrmannova, Florina Piroi, Gabriella Pasi, and Allan Hanbury. 2022. Benchmark for research theme classification of scholarly documents. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 253–262.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *ArXiv*, abs/1309.4168.

Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 11–20, Denver, Colorado. Association for Computational Linguistics.

Saif M. Mohammad. 2020. Examining citations of natural language processing literature. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5199–5209, Online. Association for Computational Linguistics.

Tsendsuren Munkhdalai, John P Lalor, and Hong Yu. 2016. Citation analysis with neural attention models. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*, pages 69–77.

Suchetha N. Kunnath, David Pride, Drahomira Herrmannova, and Petr Knoth. 2021. Overview of the 2021 SDP 3C citation context classification shared task. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 150–158, Online. Association for Computational Linguistics.

Hoa Nguyen and Steffen Eger. 2024. Is there really a citation age bias in nlp? *ArXiv*, abs/2401.03545.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125.

Alan L. Porter and Ismael Rafols. 2009. Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, 81(3):719–745.

Vinodkumar Prabhakaran, William L Hamilton, Dan McFarland, and Dan Jurafsky. 2016. Predicting the rise and fall of scientific topics from trends in their rhetorical framing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1170–1180.

Vinodkumar Prabhakaran and Owen Rambow. 2014. Predicting power relations between participants in written dialog from a single thread. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 339–344, Baltimore, Maryland. Association for Computational Linguistics.

Vahed Qazvinian and Dragomir Radev. 2008. Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 689–696.

Diana Rhoten and Andrew Parker. 2024. Risks and Rewards of an Interdisciplinary Research Path. *Science*, 306(5704):2046.

Mukund Rungta, Janvijay Singh, Saif M. Mohammad, and Diyi Yang. 2022. Geographic citation gaps in NLP research. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1371–1383, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Janvijay Singh, Mukund Rungta, Diyi Yang, and Saif Mohammad. 2023. Forgotten knowledge: Examining the citational amnesia in NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6192–6208, Toronto, Canada. Association for Computational Linguistics.

Sonita Te, Amira Barhoumi, Martin Lentschat, Frédérique Bordignon, Cyril Labbé, and François Portet. 2022. Citation context classification: Critical vs non-critical. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 49–53.

Stephen E. Toulmin. 2003. *The Uses of Argument*. Cambridge University Press.

V. Kelly Turner, Karina Benessaiah, Scott Warren, and David Iwaniec. 2015. Essential tensions in interdisciplinary scholarship: navigating challenges in affect, epistemologies, and structure in environment–society research centers. *Higher Education*, 70(4):649–665.

Richard Van Noorden. 2015. Interdisciplinary research by the numbers. *Nature*, 525:306–7.

Svitlana Volkova, Yoram Bachrach, Michael Armstrong, and Vijay Sharma. 2015. Inferring latent user properties from texts published in social media. In *AAAI Conference on Artificial Intelligence*.

Wilhelm Vosskamp, Raymond C Miller, and Julie Thompson Klein. 1986. From scientific specialization to the dialogue between the disciplines. *Issues in Interdisciplinary Studies*.

Winston Wu, Lu Wang, and Rada Mihalcea. 2023. Cross-cultural analysis of human values, morals, and biases in folk tales. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5113–5125, Singapore. Association for Computational Linguistics.

Chao Yang, Shimei Pan, Jalal Mahmud, Huahai Yang, and Padmini Srinivasan. 2015. Using personal traits for brand preference prediction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 86–96, Lisbon, Portugal. Association for Computational Linguistics.

Hao Yang, Yanyan Zhao, and Bing Qin. 2022. Face-sensitive image-to-emotional-text cross-modal translation for multimodal aspect-based sentiment analysis. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3324–3335, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dani Yogatama, Michael Heilman, Brendan O'Connor, Chris Dyer, Bryan R Routledge, and Noah A Smith. 2011. Predicting a scientific community's response to an article. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 594–604.

Dong Zhou, Séamus Lawless, Xuan Wu, Wenyu Zhao, and Jianxun Liu. 2016. Enhanced personalized search using social data. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 700–710, Austin, Texas. Association for Computational Linguistics.

A Appendix

A.1 Detailed Dataset Statistics

In Figure 12 we present the full dataset, including labelled and unlabelled data, by publication year. Figure 13 shows this dataset broken down by the conference or workshop in which it was published.

Figure 14 shows the labelled subset of data broken down by year and label. Figure 15 breaks this subset down by publication conference/workshop.

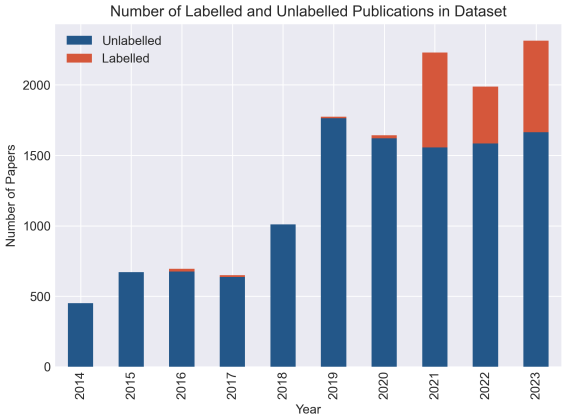

Figure 12: Full unlabelled dataset by year

A.2 Details on Citation Mapping

The Grobid and SciPDF pipeline result in a dictionary format including the article text and citations, as well as section breakdowns. To assign canonical section titles to each section, we string matched on a set of possible titles common in publications associated with our predetermined section titles. These are included in Tab. 6.

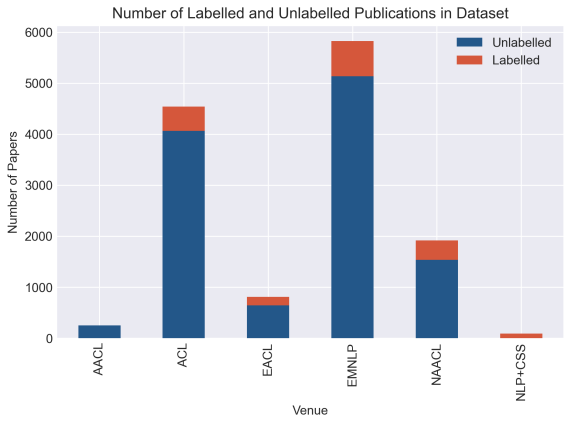

Figure 13: Full unlabelled dataset by conference

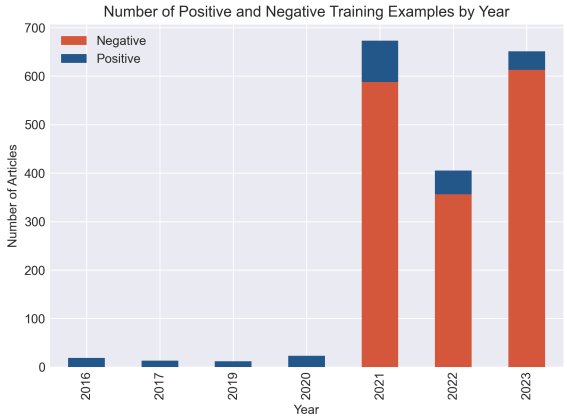

Figure 14: Labeled dataset by year

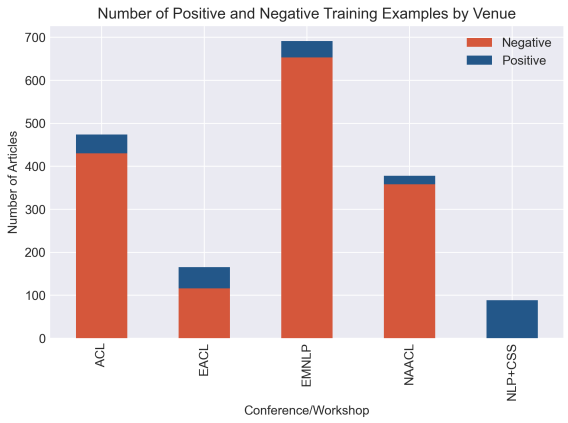

Figure 15: Labeled dataset by conference

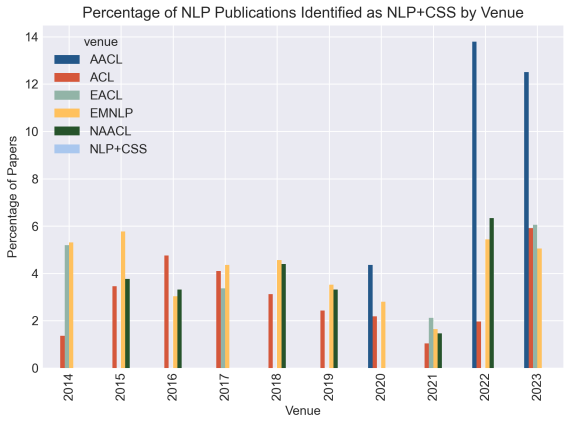

Figure 16: Percentage of CSS papers by year and venue

The dictionary includes in-text citations for every section. We matched these to an entry in the references section by string matching on first author last name and publication date.

A.3 Classifier Data Details

In Tab. 7 we show the number of examples present in each split during 5-fold cross-validation for the track classifier.

In Tab. 9 we show the number of examples present in each split during 5-fold cross-validation for the argument classifier.

A.4 Social Science Fields

accounting, anthropology, applied psychology, art, arts and humanities (miscellaneous), biological psychiatry, business, communication, criminology, cultural studies, demography, developmental and educational psychology, ecology, evolution, behavior and systematics, economics, economics and econometrics, economics, econometrics and finance (miscellaneous), education, environmental science, epidemiology, ethics, experimental and cognitive psychology, finance, gender studies, general health science, general humanities, general psychology, general social science, genetics, geography, geography, planning and development, health informatics, health information management, health policy, history, history and philosophy of science, humanities, informatics, information science, information systems, information systems and management, language and linguistics, library and information science, life-span and life-course studies, linguistics, linguistics and language, literature, marketing, pediatrics, perinatology and child health, philosophy, policy, political science, political science and international relations, press/news, psychiatry and mental health, psychology, public administration, public health, rehabilitation, social psychology, social science (miscellaneous), sociology, sociology and political science, speech and hearing, strategy and management

Canonical Title	Matching Sections
Introduction	introduction
Related Work	related work, background, related research
Method	method, methodology, approach, notation, technique, algorithm, architecture, design, solution, method, approach, methodology, model, corpus, data
Experiments	experiment, evaluation, test, analysis, compare, accuracy, scores, our method vs., state-of-the-art, baseline, results, performance, evaluation
Conclusion	conclusion, future work, discussion, limitation, ethical consideration, ethics
Appendix	appendix

Table 6: Paper section titles mapped to our canonical titles

Fold	Train	Validation	Test
0	1145	288	358
1	1145	288	358
2	1145	288	358
3	1147	286	358
4	1145	287	359

Table 7: Number of examples in each data split used for 5-fold cross-validation with task classifier

Fold	CSS	Not CSS	Macro
0	0.8872	0.9354	0.9093
1	0.9482	0.9482	0.9482
2	0.8799	0.8908	0.8852
3	0.8701	0.8916	0.8804
4	0.9031	0.8589	0.8790
Avg	0.9136	0.9362	0.9249
Stdev	0.03194	0.0158	0.0238

Table 8: F1 for CSS Track Prediction per Fold

A.5 Topic Modelling Details

We conducted topic modelling with tomotopy (Lee, 2022) over all abstracts. We modelled 15 topics and eliminated the 50 most-common words. Based on the top 10 words associated with each topic, we manually assigned the topic a title. The top ten words associated with each topic and the manually assigned label are included in Tab. 11.

Fold	Train	Validation	Test
0	5313	1537	1887
1	5608	1567	1562
2	5657	1498	1582
3	5303	1779	1655
4	5478	1208	2051

Table 9: Number of examples in each data split used for 5-fold cross-validation with argument classifier

Fold	Own Claim	Background Claim	Data	Macro
0	0.7989	0.6508	0.5	0.6499
1	0.7031	0.6251	0.4367	0.5883
2	0.7652	0.5957	0.4389	0.5999
3	0.763	0.6405	0.5059	0.6365
4	0.7516	0.597	0.491	0.6132
Avg	0.7564	0.6218	0.4745	0.6176
Stdev	0.0346	0.0250	0.0339	0.0255

Table 10: F1 for Argument Type Prediction per Fold

Topic Name	Top 10 Words
Discourse	context, only, but, at, time, also, discourse, study, online, evidence
Sarcasm	al, sarcasm, et, moral, its, aspect, have, methods, been, then
State of the Art	over, approach, stateoftheart, baseline, present, outperforms, stance, novel, based, annotated
Emotion Detection	emotion, knowledge, information, datasets, demonstrate, experiments, multimodal, stateoftheart, effectiveness, rumor
Twitter Data	users, user, twitter, tweets, posts, predict, emotional, methods, individuals, studies
Mental Health	learning, framework, existing, health, mental, large, tasks, natural, novel, proposed
Gender	more, study, gender, than, nlp, have, may, find, groups, people
Political Issues	political, how, identify, us, computational, articles, through, identifying, science, issues
Hate Speech	speech, content, hate, not, online, detecting, one, also, but, personal
Semantic Structure	semantic, used, information, annotated, human, structure, them, documents, sentences, set
Linguistics	features, new, research, languages, linguistic, into, english, how, across, while
Word Embeddings	approach, words, word, embeddings, use, method, all, same, two, predicting
Events	eg, information, about, prediction, both, events, methods, also, change, event
Machine Learning	training, bias, classification, datasets, trained, both, at, through, been, problem
Conversations	conversations, conversation, strategies, computational, power, persuasion, where, framework, not, conversational

Table 11: 15 topics identified in abstracts with hand-labelled titles