How Do We Engage with Other Disciplines? A Framework to Study Meaningful Interdisciplinary Discourse in Scholarly Publications

Bagyasree Sudharsan Alexandria Leto Maria Leonor Pacheco
University of Colorado Boulder
{bagyasree.sudharsan, alexandria.leto, maria.pacheco}@colorado.edu

Abstract

With the rising popularity of interdisciplinary work, there is a growing need to understand how publications incorporate ideas from multiple disciplines. Existing computational approaches such as affiliation diversity, keyword matching, and citation patterns do not account for *how* individual citations are used to advance the citing work. Prior studies have proposed taxonomies to classify citation purpose. However, these frameworks are not well-suited to interdisciplinary research and do not provide general measures of engagement *depth*. To address these limitations, we propose a framework for evaluating citation engagement in scholarly publications. We introduce a citation purpose taxonomy tailored to interdisciplinary work, supported by an annotation study and a model for automatically predicting a citation's purpose. We then leverage signal from a citation's purpose and the sections in which it appears to predict a single measure of how deeply a publication engages with a citation. We demonstrate the utility of our framework through a large-scale analysis of publications at the intersection of NLP and Computational Social Science.

1 Introduction

Interdisciplinary research involves the integration of concepts, perspectives, theories, methods, and data from two or more bodies of specialized knowledge to construct creative solutions (National Academy of Sciences et al., 2005). This definition highlights knowledge integration as a central component of interdisciplinarity, suggesting that it should be assessed not only in terms of the amount of knowledge integrated but also how it is combined. Meta-analyses which seek to understand the convergence of, and interplay between, different fields should therefore aim to capture the *depth* with which an article engages with the work it cites.

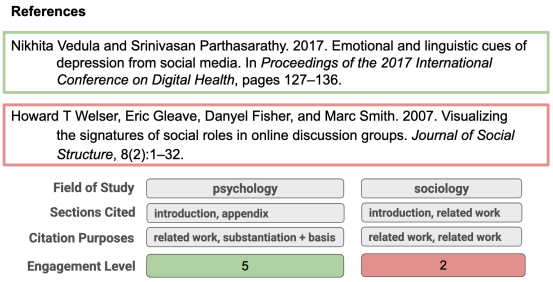

Figure 1: Framework for evaluating a publication's level of engagement with citing work based on the section and purpose of all in-text references.

Contrary to this, much of the existing meta-analytic work has focused on broad, scalable indicators (e.g., diversity and entropy of citations) (Porter and Rafols, 2009; Van Noorden, 2015; Cassi et al., 2017; Zhao and Yang, 2023) that emphasize the quantity and diversity of knowledge sources, but offer limited insight into how deeply or substantively those sources are integrated. For example, an article may cite a broad set of disciplines in its *Introduction* or *Related Work* sections without explicitly building on the cited work's ideas or using them to inform methodological choices.

Motivated by the need to go beyond traditional measures of interdisciplinarity, we propose a framework for characterizing the depth with which an article engages with the works it cites.[1] We build on the framework by Leto et al. (2024), which is motivated by the intuition that the section in which a citation appears signals different levels of engagement. It extracts in-text citations from interdisciplinary papers and maps them to the section in which they occur and identifies whether they are out-of-discipline from the citing work. We extend this framework by introducing a novel taxonomy to represent the purpose for which an author includes a given citation. Although numerous citation-purpose classification schemes have been

[1] Code, models, and data.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 49070–49090
October 24-29, 2026 ©2026 Association for Computational Linguistics

proposed (Jurgens et al., 2018; Abu-Jbara et al., 2013; Lauscher et al., 2022; Cohan et al., 2019; Teufel et al., 2006; Pride and Knoth, 2020), existing taxonomies are insufficient for analyzing interdisciplinary citations, either because they lack relevant categories, include categories that are ill-suited to cross-disciplinary contexts, or fail to incorporate adequate contextual information.

Our taxonomy, shown in Figure 1, addresses these limitations by situating each citation within its local argumentative context and characterizing citation purpose through categories designed to reflect the different forms of interdisciplinary engagement that arise when authors draw on work beyond their primary field. The taxonomy was developed from the ground up through an inductive annotation study, allowing categories to emerge from close analysis of real citation practices rather than being imposed a priori.

We operationalize our framework, integrating signal from a citation's purpose and the section in which it appears into a single **Citation Engagement Predictor (CEP)**. Our CEP provides a measure of the depth of a publication's engagement with a work it cites, allowing practitioners to gauge engagement at a glance. We ground our proposed CEP in an annotation study and statistical analysis. These analyses find that high-engagement citations are rare and some purposes are more common in interdisciplinary works. Together with our qualitative analysis, they suggest that a dedicated effort towards making citations more integrative would result in better interdisciplinary work.

Our framework, annotation studies, and analysis are developed in the context of Natural Language Processing (NLP) research, positioning these contributions as a first step toward a broader framework for studying interdisciplinary engagement across fields. We focus specifically on work at the intersection of NLP and Computational Social Science (CSS). Although research in this area has grown substantially over the past decade (Grimmer and Stewart, 2013), the quality of NLP+CSS publications has been questioned, with critiques arguing that such work often falls short of social science standards of rigor (Baden et al., 2022) and makes limited substantive contributions to those fields (McCarthy and Dore, 2023). We use our framework to further examine these concerns and to evaluate the extent to which citation engagement can be reliably studied at scale, including through the use of automated methods. We make the following contributions:

1. A novel annotation framework designed to capture citation purpose in interdisciplinary NLP work. We accompany this with a dataset of 950 annotated in-text citations from NLP+CSS papers and an automatic classifier for predicting a citation's purpose according to our taxonomy.

2. A Citation Engagement Predictor designed to provide a rating of a publication's engagement with a given cited work based on signal from the purpose and section of corresponding in-text references.

3. A case study in NLP+CSS work demonstrating the usefulness of our framework. We find that articles with many high-engagement interdisciplinary citations are more principled and better motivated, while many within-discipline citations provide more rigorous dataset analysis. Interdisciplinary citations are more often engaged with at the idea- and claim-level, while within-discipline citations drive methodology.

2 Related Work

Numerous studies have introduced frameworks for classifying citation purpose or intent, but these approaches are not designed to capture how authors engage with references from outside their primary discipline. For example, Jurgens et al. (2018) and Pride and Knoth (2020) introduce datasets that are heavily dominated by a *Background* category designed to denote when a reference provides relevant information for the domain of the article. By treating all background citations as equivalent, this definition fails to distinguish between references that provide foundational knowledge necessary for interpreting the work and those that merely provide general or loosely related context. Similarly, their *Extension* category is narrowly defined to apply only when authors directly extend a method or dataset; a scenario that is relatively rare in interdisciplinary contexts, where engagement often takes other forms. In addition, their framework relies on a fixed citation context of approximately 300 surrounding characters, which may be insufficient to capture the broader argumentative role needed to accurately determine citation purpose. Lauscher et al. (2022) and Abu-Jbara et al. (2013) address this limitation by considering a window of

sentences surrounding each citation — an approach that we adopt in this work.

Teufel et al. (2006) propose a highly fine-grained scheme for classifying citation function, with multiple sub-categories that distinguish forms of comparison and contrast. While this level of detail can be valuable, the resulting categories primarily capture evaluative stance and intra-discipline relations such as direct comparison and contrast. Additionally, their dataset labels over 60% of citations as *Neutral*, limiting its ability to distinguish meaningful forms of engagement. In contrast, our scheme places only $\sim$ 38.5% of citations into similar general categories (see Table 2), allowing for finer differentiation among types of interdisciplinary use.

At the opposite side of the spectrum, Cohan et al. (2019) introduce a taxonomy with only three high-level categories, which are too coarse-grained to adequately capture variation in the depth of citation engagement required for our analysis.

3 Data

In this section, we discuss our methods for collecting and processing our dataset of interdisciplinary NLP+CSS papers.

3.1 Interdisciplinary Article Collection

We construct a dataset of CSS papers from the ACL Anthology[2] published between 2014-2025. To do so, we build on a pre-existing data collection methodology (Leto et al., 2024), where an automatic track classifier is used to identify NLP+CSS publications. Approximately 1,800 abstracts were labeled with their track using conference programs from 2021-2023. Papers categorized under "Computational Social Science and Cultural Analytics" serve as positive examples, while negative examples are drawn from all other tracks. We augment the positive set with papers from the ACL Special Interest Group on Language Technologies for the Socio-Economic Sciences and Humanities (SIGHUM) workshop and the Natural Language Processing and Computational Social Science (NLP+CSS) workshop. The resulting dataset comprises 497 positive and 2,692 negative examples, for a total of 3,189 labeled abstracts. The distribution of labels over time is shown in Figure 8. We use this labeled dataset to train a binary classifier to identify CSS papers. Specifically, we fine-tune RoBERTa-base (Gururangan et al., 2020) on abstract text. Details are provided in Appendix A.

We download all long publications in the ACL anthology from 2014-2025. Then, we use the classifier (details in Appendix A.1) to identify 1,534 of the 17,258 unlabeled articles as NLP+CSS publications. Combined with the initial positively-labeled articles, this results in a dataset of 2,031 NLP+CSS articles.

3.2 Citation Mapping

After the CSS papers are identified, their content is processed using Grobid[3], then converted into dictionary representation using SciPDF Parser[4]. The resulting dictionary contains a breakdown of the textual content, including the reference section. Section names are mapped to canonical sections *Introduction*, *Related Work*, *Method*, *Experiments*, *Conclusion*, *References*, and *Appendix* (details in Appendix A.2). The in-text references are then extracted from each section and assigned to their entry in the *References* section, connecting them to the title of the cited publication, the publication venue and other publication meta-data. Finally, cited works are mapped to a "field of study" by querying the Semantic Scholar API (Kinney et al., 2023) or using a string-matching approach. All citations with a field of study outside of "Computer Science" and "Linguistics" are determined to be out-of-discipline.

Our resulting dataset of NLP+CSS work contains 2,031 articles (see Figure 2). From these, we extracted 50,050 in-text references to 19,051 unique cited works. 20,319 of these in-text references (17,996 cited works) were determined to be out-of-discipline, for an average of $\sim$ 10 out-of-discipline in-text citations per article.

4 Citation Purpose

In this section, we discuss the novel taxonomy developed to classify citation purpose, the corresponding annotation process for obtaining a manually-labeled dataset, and an automatic classifier for citation purpose prediction.

4.1 Citation Purpose Taxonomy

Measuring the level of interdisciplinary engagement in a publication requires identifying the role that out-of-discipline references play in shaping the paper's conceptual, methodological, and empirical

[2] https://aclanthology.org/

[3] https://github.com/kermitt2/grobid

[4] https://github.com/titipata/scipdf_parser

Category	Explanation
SUBSTANTIATION +BASIS	The citation both substantiates a claim made in the citing work and serves as the basis for an idea or method that is subsequently built upon.
BASIS	The method, idea, or tool in the citation serves as the basis for the citing work. This must be explicitly stated or clearly evident from the abstract that the cited idea is central and explored further.
SUBSTANTIATION	The citation is used to verify or substantiate theoretical, empirical, or methodological claims made in the citing work. When classified as Substantiation, the supported claim should be identifiable.
USE	The method or tool in the citation is used directly in the citing work, with no or very trivial modifications.
DEFINITION	The citation is used to define or explain a topic or phrase (but not to validate it).
ANALYSIS	The method or idea in the citation is analyzed through comparisons or critiques that are directly linked to the citing work. The cited work may be compared to or criticized to justify an approach taken in the citing work, but is not explicitly built upon.
RELATED WORK	The citation refers to work that is related to the citing work or provides it as an example of a method, idea, or phenomenon discussed without further description. It does not fit any of the other categories.

Table 1: Citation purpose categories used for annotation.

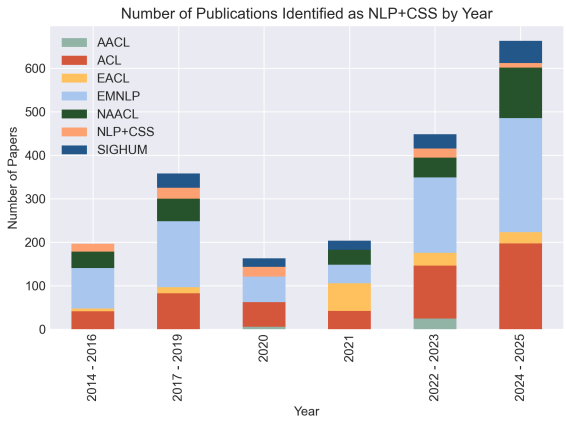

Figure 2: Dataset of NLP+CSS articles across years and venues

claims. For instance, a citation that contrasts a proposed approach with prior methods in the literature does not necessarily reflect deep engagement. In contrast, citations that reference methods or claims that are foundational to the analyses or arguments advanced in the published work indicate a substantially higher degree of engagement.

In line with this goal, we propose a citation-purpose classification scheme adapted from Abu-Jbara et al. (2013), who define six categories: *Criticizing*, *Comparison*, *Use*, *Substantiating*, *Basis*, and *Neutral/Other*. Through an iterative, inductive annotation process conducted by two of the authors of the paper, we refine and extend this framework to better capture the roles that citations play in interdisciplinary contexts and to support assessments of engagement depth. The resulting taxonomy is presented in Table 1. Label determinations are made considering the paragraph surrounding each citation and the claims made in the paper's abstract, which together provide sufficient argumentative context for labeling decisions.

The categories *Basis* and *Substantiation + Basis* point to a deeper engagement with the literature. The latter, in particular, is designed to highlight deep engagement with interdisciplinary work. Often a claim is validated by an out-of-discipline citation, then used to inform the proposed method. For example, in the sentence

> "As pronoun usage tends to differ in individuals living with depression **(Vedula and Parthasarathy, 2017)**, we removed any English pronouns from our stop word set"

a citation supporting a claim from the field of psychology directly informs a methodological choice.

Another key feature of our taxonomy is the inclusion of the *Analysis* category, which combines the frequently-used categories *Comparison* and *Criticism* (Abu-Jbara et al., 2013). In most cases, these categories refer to the same general principle: analyzing the idea in the citation. Whether the ideas are being compared to or criticized does not signal different depths of engagement. The introduction of the *Definition* category helps identify citations that establish core concepts used in the work. Although citing prior work to support a definition may appear surface-level, such citations signal the conceptual importance of the defined term within the work and therefore provide meaningful evidence about the depth of interdisciplinary engagement.

Category	Interdisciplinary	Within Disc.	Total
Related Work	163	202	**365**
Substantiation	98	86	**184**
Use	54	115	**169**
Analysis	31	77	**108**
Basis	18	26	**44**
Substantiation + Basis	26	17	**43**
Definition	23	14	**37**

Table 2: Number of annotated citations per citation purpose category with full or partial agreement.

Finally, we specifically design the *Related Work* category to include citations which support claims about the state of the literature (e.g. "There exist studies on this subject", "Multiple works have adopted this approach"). This distinction separates such citations from those that support empirical or findings-based claims, which are more appropriately categorized as *Substantiation*. In interdisciplinary work, the latter typically reflects a deeper and more consequential form of engagement.

4.2 Citation Purpose Annotation

To create our dataset, we performed an annotation study with three of the paper authors and two computer science graduate students. The graduate students received training from the authors and were provided with our Codebook (Appendix G).

We selected 248 articles at random, balanced across publication year and venue. We randomly selected $\sim 1,000$ in-text citations from this set, balancing across the section in which they appear. 624 of the 998 (62.5%) examples were doubly-annotated, and 374 (37.5%) were triply-annotated. Agreement statistics were calculated over the entire annotated set (both doubly- and triply-annotated). We had complete agreement on 646 of the citations and partial agreement on 304 (at least partial agreement on 950). We obtained a Krippendorff's α of 0.596.

The resulting dataset of 950 citations (with at least partial agreement) is summarized in Table 2. A pairwise comparison of annotations is included in Figure 3. A comprehensive discussion of common disagreements is included in Appendix D.

4.3 Automatic Citation Purpose Classification

Here, we present and evaluate SOTA models for automatically classifying citation purpose according to our citation purpose taxonomy. We start by splitting our annotated dataset, at the article level, into a training set consisting of 80% of the data and a held-out test set of the other 20% of the data. Though our classifier operates at the in-text citation level, splitting by article guarantees no within-article leakage. We evaluate three models: (1) fine-tuned RoBERTa, (2) prompting `Llama3-8b` (Grattafiori et al., 2024), and prompting `Qwen3-8b` (Yang et al., 2025).

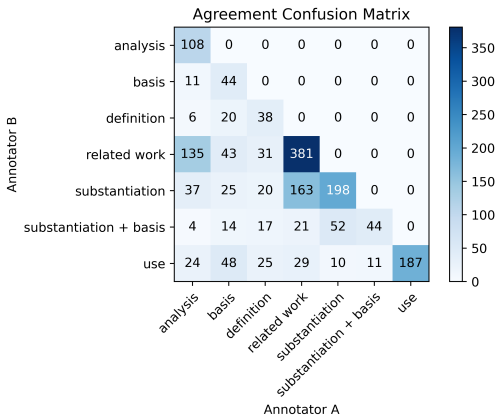

Figure 3: Pairwise comparison of annotations. Each annotation for a sample was considered with each of the other annotations (i.e., there is one pair for doubly-annotated samples and three pairs for triply-annotated samples). Annotator IDs on the x- and y-axes are arbitrary.

Model	Precision	Recall	Weighted F1	Macro F1
Random	0.18	0.18	0.22	0.16
`Llama3`	0.30	0.25	0.19	0.15
`Qwen3`	**0.58**	**0.58**	**0.60**	**0.56**
`RoBERTa`	0.48	0.50	0.60	0.49

Table 3: Automatic citation purpose classification results.

We fine-tuned RoBERTa on the training set using the AdamW optimizer, weighted cross-entropy loss, and a learning rate of $4e-5$. For early stopping, we use the macro F1 on the validation set (10% of the training set). Temperature was set to 1 for `Llama3-8b` and `Qwen3-8b`; prompts are provided in Appendix H. We report precision, recall, weighted f1, and macro f1 performance on the held-out test set in Table 3. We include detailed classification reports for each model in Appendix B.

Results We find that `Qwen3` out-performs all other models, achieving the greatest precision, recall, weighted f1-score, and macro f1-score. We use our classifier to automatically obtain Citation Purpose classifications across the entire unlabeled dataset used for the case study in Section 6.

5 Citation Engagement Predictor

We posit that the depth of engagement reflected by a citation is captured by two components: the purpose and the section of its in-text references. In this section we present a Citation Engagement Predictor (CEP) which seeks to combine these components for a singular rating of a publication's engagement with a given citation. We define a 5-point scale for this rating and present a method for predicting it automatically.

5.1 Citation Engagement Annotation Study

We run an annotation study to ground our CEP in expert judgments. We randomly sample 100 article-citation pairs, distributed as evenly as possible by frequency of occurrence of the citation in the article and by article year and venue, to be representative of the full dataset. Two computer science researchers with expertise in NLP+CSS work analyzed how each cited work was used throughout the article, considering all of its in-text references, and assigned it an engagement score on a five-point Likert scale. A score of one indicates that the annotator "strongly disagrees" with the statement "The article engages deeply with the cited work", and five indicates "strongly agrees". On a first pass, we achieved high inter-annotator agreement (ordinal Krippendorff's α of 0.84). The thirty resulting disagreements tended to occur with only a one-point difference. The annotators deliberated on disagreements to arrive at a final set of annotations. For each of these 100 cited works, we also had the same graduate students hand-annotate the purpose of each citation occurrence, and obtained section labels from our data collection pipeline (Section 3).

5.2 Modeling Engagement

For each cited work, we construct features representing how often it is referenced with each purpose and in each section. We fit an ordinal logistic regression model on our 100 annotated cited works to learn the boundaries between engagement levels and the contribution of each feature to engagement. We use 7 purpose category features and 6 section categories for a total of 13 features.

Because every in-text reference is assigned exactly one purpose label and appears in exactly one section, the 7 purpose counts and the 6 section counts for each cited work both sum to its total number of in-text references. This creates a linear dependency between the purpose and section features, which prevents the model from uniquely estimating their coefficients. To address this, we drop one category from each dimension as the baseline. We choose "Related Work" in both cases, since it intuitively signals the least engagement and keeps the remaining coefficients positive and easier to interpret. This leaves us with 11 features. We set the value of each feature to $log(1 + n_x)$, where n_x is the number of in-text references with that purpose or in that section.

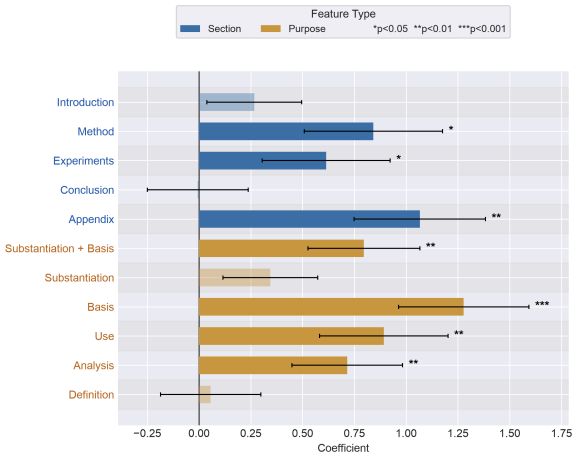

Figure 4: Regression coefficients per feature.

We perform a 90-10 train-test split on our set of 100 article-citation pairs, at the article level, to ensure that the results were not skewed by within-article leakage. A 10-fold cross validation on our model showed low variance and standard deviation across folds, providing confidence that the model is reliable and generalizable.

5.3 Results

We achieve a Mean Absolute Error (MAE) of $0.476^{\pm 0.110}$, a Quadratic Weighted Kappa (QWK) of $0.813^{\pm 0.064}$, and an accuracy of $0.563^{\pm 0.129}$, averaged across the 10 folds. We report coefficients in Figure 4, which tell us the increase in log-odds of being rated at a higher engagement level as the citation counts per section or purpose increase. We see that having the purpose *Basis* and being in the *Appendix* section are especially indicative of a citation being engaged with highly, with low p-values. Notably, the p-values of the coefficients for the *Definition* purpose category and the *Conclusion* section are particularly high, suggesting that more data is required to properly draw conclusions about the effects of these features on engagement levels. In general, we find that the purpose of a citation,

Features	MAE	QWK
All	0.476 ± 0.110	0.813 ± 0.064
Section only	0.585 ± 0.070	0.749 ± 0.108
Purpose only	0.556 ± 0.116	0.716 ± 0.112

Table 4: Model performance with only a subset of features.

Features Dropped	LR	p-value
Purpose	30.3	0.00003
Section	16.4	0.005

Table 5: Log-likelihood ratios with features dropped in comparison to the full feature set.

i.e., why a citation is used, offers more insight into its level of engagement in an article than where it occurs. This finding underscores the value of our extension to Leto et al. (2024). While section alone carries some signal, the purpose of a citation (as captured by our taxonomy) is the more informative component, and incorporating it yields a substantially richer measure of engagement.

We conduct ablation studies to further confirm these findings and validate our choice to use both sets of features. Table 4 shows the MAE and QWK when only a subset of features is used. The lower performance in both cases indicates that both the section and purpose feature sets are necessary and informative. Table 5 shows the log-likelihood ratios with features dropped. The high ratios with respect to the full set indicate that the improvement from the addition of features is not just chance and arises from a substantial improvement in predictive power. It is also clear from the latter experiment that purpose is the stronger predictor, as removing purpose features results in a substantially larger drop in log-likelihood.

6 NLP + CSS Engagement Analysis

In this section, we analyze our dataset to demonstrate the utility of our proposed Citation Purpose framework and Citation Engagement Predictor. We leverage the automatic Citation Purpose classifier (Section 4) to obtain citation purpose classifications for all in-text citations in our dataset. Then, we use our Citation Engagement Predictor to obtain the level of engagement (Section 5) for each citation-article pair in our dataset, considering all occurrences of the citation within the article, and their purposes and sections. We drop pairs which

Level	Mean	Median	Maximum
1	11.05	9	78
2	4.6	4	44
3	1.19	1	16
4	0.22	0	6
5	0.18	0	5

Table 6: Number of citations per engagement level per article.

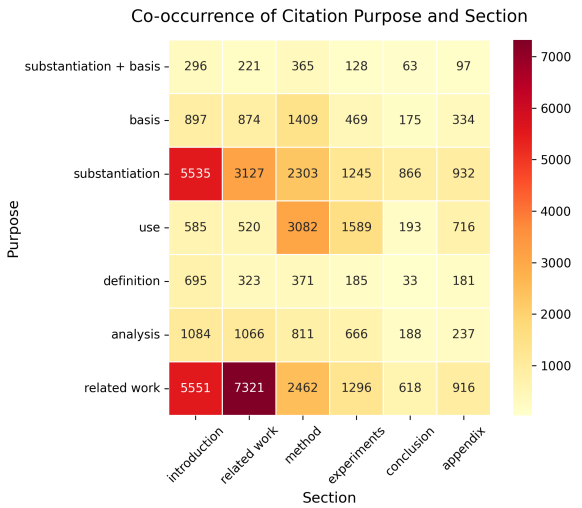

Figure 5: Co-occurrence of citation purpose and section in full dataset.

do not have purpose or section information for all occurrences. While individual purpose predictions and engagement estimates carry some error, we expect aggregate trends across our dataset of 34956 article-citation pairs to be informative. We present these analyses below.

6.1 Quantitative Analysis

Articles have more low-engagement citations. Out of all article-citation pairs, we find that 64.1% (22398) have Level 1 (low) engagement, 26.7% (9328) and 6.9% (2416) have Levels 2 and 3 (medium) engagement, and only 1.3% (443) and 1.1% (371) have Levels 4 and 5 (high) engagement. We present statistics on the number of citations per engagement level per article in Table 6, finding that papers tend to engage with only a few citations on a deeper level.

Some sections and citation purposes intuitively co-occur. To predict the engagement level, we use the purposes of the citation in the article and the sections of occurrence; the co-occurrence between these features on our complete dataset, at the in-text citation level, is in Figure 5. We find that many citations used for the purpose of *Related*

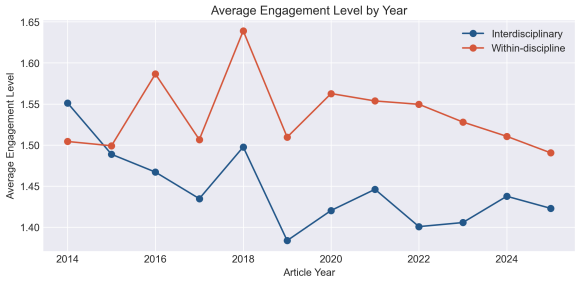

Figure 6: Average engagement level by article year.

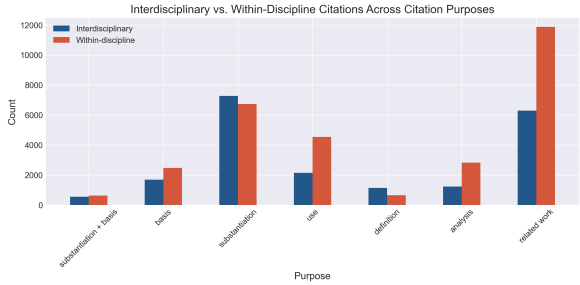

Figure 7: Interdisciplinary vs. Within-Discipline citations by citation purpose.

Work often appear in the *Introduction* and *Related Work* sections. *Substantiation* also often occurs in the *Introduction*, while *Use*-purpose citations are common in the *Method* and *Experiments* sections. However, we see enough variation in these patterns to require both components to predict engagement. We verify that these trends hold in the subset of citations hand-annotated with citation purpose (see Appendix Figure 11).

Articles show deeper engagement with within-discipline citations. In Figure 6, we show the change in average engagement level by year. We find that authors tend to engage more deeply with within-discipline citations than interdisciplinary citations. However, this trend does not hold for the year 2014, although the difference is minimal.

Certain purposes are more associated with interdisciplinary citations. An analysis of the purpose counts of within-discipline vs. out-of-discipline citations is in Figure 7. This plot shows that the categories *Substantiation* and *Definition* in particular are more frequent for interdisciplinary references. This is also reflected in the subset of citations hand-annotated with citation purpose (see Appendix Figure 12).

6.2 Qualitative Analysis

We conduct a qualitative analysis of a set of articles to understand how various engagement levels with cited works manifest in a complete paper (see Table 7). We select two articles that have 4+ high-engagement citations (levels 4 and 5) — one where the majority of these are interdisciplinary, and one where they are not. We compare these to articles which have only 1-2 high-engagement citations, as well as articles which have no high-engagement citations.

The article "When Do Word Embeddings Accurately Reflect Surveys on our Beliefs About People?" (Joseph and Morgan, 2020), has 4 high-engagement citations. They engage with multiple citations ('Kozlowski et al., 2019', 'Swinger et al., 2019'), to motivate their design of embedding concepts to recover beliefs, making their adoption of concepts from another discipline — cognitive science — principled and well-motivated. Their experiments on specific belief taxonomies proposed by other high-engagement citations ('Smith-Lovin and Robinson, 2015'), add further rigor to their contributions.

The article "Personality Matters: User Traits Predict LLM Preferences in Multi-Turn Collaborative Tasks" (Yunusov et al., 2025) similarly engages with cognitive traits, but our framework assigns it lower engagement scores across its citations, interdisciplinary and otherwise. The authors adopt a taxonomy from the cognitive science literature, but the citation context does not extensively motivate why this taxonomy is the right tool for their analysis. In contrast, Joseph and Morgan (2020) examine the different methods of constructing dimensions, across sociocultural themes, before selecting a set; this provides additional grounding for its methodological choices.

Similarly, "Ideology Prediction from Scarce and Biased Supervision: Learn to Disregard the 'What' and Focus on the 'How'!" (Chen et al., 2023) has a small number of high-engagement citations and tends to anchor one or two central aspects of its work in them. While the article offers a thoughtful justification for this foundation, the depth of citation engagement is concentrated in those few aspects rather than distributed across the analysis, as is more common in articles with multiple high-engagement citations.

"Modeling and Prediction of Online Product Review Helpfulness: A Survey" (Ocampo Diaz and Ng, 2018) provides an example of an article that has many high-engagement, within-discipline citations. In this article, we observe that the high- and

Paper	High-Eng. Citations	Majority Interdisc.?
"When do Word Embeddings Accurately Reflect Surveys on our Beliefs About People?" (Joseph and Morgan, 2020)	4	Yes
"Modeling and Prediction of Online Product Review Helpfulness: A Survey" (Ocampo Diaz and Ng, 2018)	8	No
"Ideology Prediction from Scarce and Biased Supervision: Learn to Disregard the 'What' and Focus on the 'How'!" (Chen et al., 2023)	2	Yes
"Personality Matters: User Traits Predict LLM Preferences in Multi-Turn Collaborative Tasks" (Yunusov et al., 2025)	0	N/A
"Deciphering Oracle Bone Language with Diffusion Models" (Guan et al., 2024)	0	N/A

Table 7: Articles used for qualitative analysis.

medium-engagement citations ('Tang et al., 2013', 'Yang et al., 2015') describe datasets that are central to its work, and are engaged with in multiple ways. In contrast, the article "Deciphering Oracle Bone Language with Diffusion Models" (Guan et al., 2024), despite being closely tied to a historical writing system, does not deeply engage with any of the related work from that field, choosing instead to focus on the methodological improvement within the discipline of Computer Science.

In general, we find that interdisciplinary citations tend to be engaged with more at the idea and claim level, whereas within-discipline citations are engaged with methodologically. We see a similar pattern when comparing articles with many high-engagement citations to those with only one or two. We also find that high-engagement citations tend to span multiple sections and purposes, reflecting how deeply integrated work is woven throughout a paper. These patterns are not captured by conventional measures such as citation entropy or diversity indices, which treat all citations as equivalent. By capturing *how* authors engage with the work they cite, rather than only *what* they cite, our framework provides a complementary view to existing measures of interdisciplinarity.

7 Conclusion

Motivated by the need to evaluate the engagement of interdisciplinary publications, we present a novel annotation taxonomy designed to capture citation purpose. The taxonomy includes categories specifically designed to capture interdisciplinary citation purpose. We present an annotation study of NLP+CSS work and a resulting dataset of 950 agreed-upon annotated citations, along with an automatic classifier for predicting citation purpose. We operationalize our Citation Purpose taxonomy to present a Citation Engagement Predictor for estimating a publication's depth of engagement with a given citation given signal from the section and purpose of corresponding in-text references. Finally, we demonstrate the usefulness of our resulting framework by presenting an analysis of a set of NLP+CSS publications, finding that only a minority engage very deeply with the works they cite. Because we develop and analyze this framework in the context of NLP work, it serves as a first step for the development of a broader framework to study interdisciplinarity. We leave the validation of this framework in additional fields to future work.

Limitations

We offer an overview of the failure points of our framework in Appendix F. However, we would like to emphasize a few limitations here: (1) An automated framework was used to construct the dataset, potentially resulting in parsing errors and inaccuracies when extracting elements from scientific publications. (2) The presented analysis spans a limited number of articles and citations and includes only NLP+CSS work. Further analysis is required to understand whether the takeaways generalize to a larger dataset or additional fields of study. (3) An automatic classifier was used to predict the Citation Purpose labels used for the case study in Section 6.

As a result, the analysis carries some uncertainty.

Acknowledgments

This work utilized the Blanca high performance computing resources at the University of Colorado Boulder. Blanca is jointly funded by computing users and the University of Colorado Boulder.

References

Amjad Abu-Jbara, Jefferson Ezra, and Dragomir Radev. 2013. Purpose and polarity of citation: Towards NLP-based bibliometrics. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 596–606, Atlanta, Georgia.

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.

Lorenzo Cassi, River Champeimont, Wilfriedo Mescheba, and Élisabeth de Turckheim. 2017. Analysing institutions interdisciplinarity by extensive use of rao-stirling diversity index. *PloS*.

Chen Chen, Dylan Walker, and Venkatesh Saligrama. 2023. Ideology prediction from scarce and biased supervision: Learn to disregard the "what" and focus on the "how"! In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9529–9549, Toronto, Canada. Association for Computational Linguistics.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. 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 3586–3596, Minneapolis, Minnesota.

Shrey Desai, Cornelia Caragea, and Junyi Jessy Li. 2020. Detecting perceived emotions in hurricane disasters. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5290–5305, Online. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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.

Haisu Guan, Huanxin Yang, Xinyu Wang, Shengwei Han, Yongge Liu, Lianwen Jin, Xiang Bai, and Yuliang Liu. 2024. Deciphering oracle bone language with diffusion models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15554–15567, Bangkok, Thailand. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5427–5444, Online. Association for Computational Linguistics.

Kenneth Joseph and Jonathan Morgan. 2020. When do word embeddings accurately reflect surveys on our beliefs about people? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4392–4415, Online. Association for Computational Linguistics.

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 C. Murray, Christopher Newell, Smita Rao, Shaurya Rohatgi, Paul Sayre, Shannon 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.

Anne Lauscher, Brandon Ko, Bailey Kuehl, Sophie Johnson, Arman Cohan, David Jurgens, and Kyle Lo. 2022. MultiCite: Modeling realistic citations requires moving beyond the single-sentence single-label setting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1875–1889, Seattle, United States.

Alexandria Leto, Shamik Roy, Alexander Hoyle, Daniel Acuna, and Maria Leonor Pacheco. 2024. A first step towards measuring interdisciplinary engagement in scientific publications: A case study on NLP + CSS research. In *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS 2024)*, pages 144–158, Mexico City, Mexico. ACL.

Sarah Masud, Sahajpreet Singh, Viktor Hangya, Alexander Fraser, and Tanmoy Chakraborty. 2024. Hate personified: Investigating the role of LLMs in content moderation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15847–15863, Miami, Florida, USA. 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.

National Academy of Sciences, National Academy of Engineering, and Institute of Medicine. 2005. *Facilitating Interdisciplinary Research*. The National Academies Press, Washington, DC.

Gerardo Ocampo Diaz and Vincent Ng. 2018. Modeling and prediction of online product review helpfulness: A survey. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 698–708, Melbourne, Australia. Association for Computational Linguistics.

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.

David Pride and Petr Knoth. 2020. An authoritative approach to citation classification. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, page 337–340, New York, NY, USA. Association for Computing Machinery.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 103–110, Sydney, Australia.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

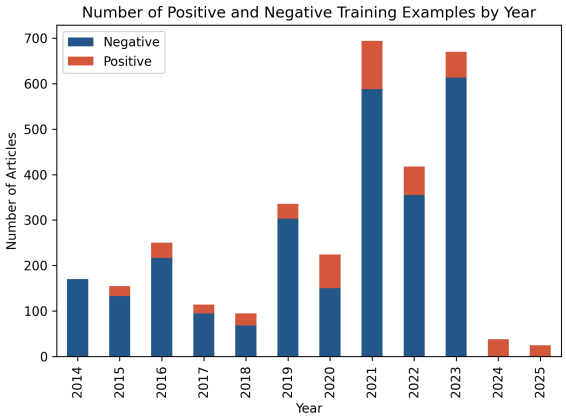

Figure 8: Labeled dataset for training track classifier

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Sarfaroz Yunusov, Kaige Chen, Kazi Nishat Anwar, and Ali Emami. 2025. Personality matters: User traits predict LLM preferences in multi-turn collaborative tasks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1359–1372, Suzhou, China. Association for Computational Linguistics.

Qun Zhao and Menghui Yang. 2023. An efficient entropy of sum approach for measuring diversity and interdisciplinarity. *Journal of Informetrics*, 17(3):101425.

A Additional Data Collection Details

Here we include additional details for our interdisciplinary article collection and citation mapping processes.

A.1 Automatic Track Classifier Details

To build our binary track classifier, we fine-tune RoBERTa base (Gururangan et al., 2020) on the labeled abstracts (Figure 8) using HuggingFace's Transformers Trainer (Wolf et al., 2020), employing a weighted cross-entropy loss to address class imbalance. The model is trained for 50 epochs

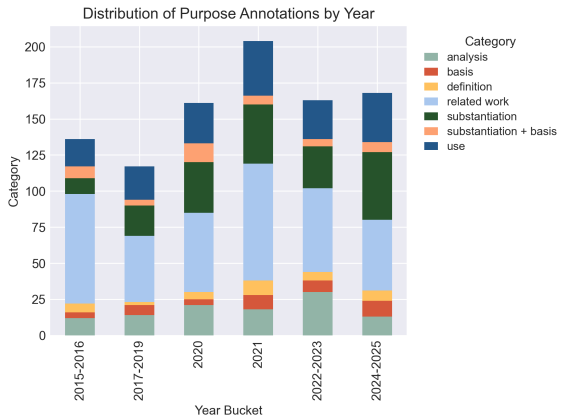

Figure 9: Distribution of agreed-upon citation purpose annotations by article year.

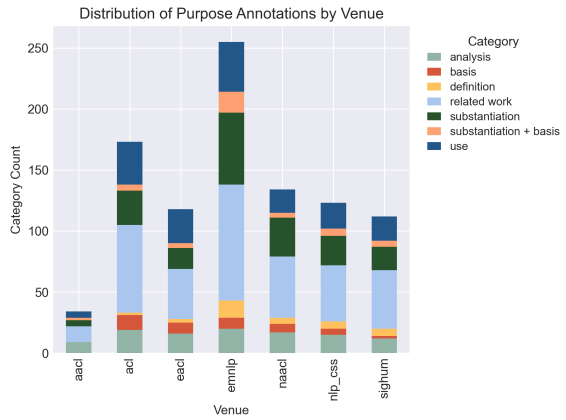

Figure 10: Distribution of agreed-upon citation purpose annotations by article venue.

with a learning rate of $2e-5$, using an 80/10/10 train/validation/test split. The checkpoint achieving the highest macro F1 score on the validation set is selected, yielding a macro F1 of 0.870, precision of 0.863 and recall of 0.876 on the test set.

A.2 Mapping In-text References to Sections

To arrive at the canonical section, the section titles initially parsed by Grobid are assigned to one of "Introduction," "Related Work," "Method," "Experiments," "Conclusion," or "Appendix" using string-matching rules. The first section is always mapped to "Introduction." Then, additional sections are mapped to titles if they contain the keywords shown in Table 8. Sections that cannot be mapped are not used for analysis.

B Citation Purpose Classifier Details

In this section we include full classification reports for `Llama3` in Table 9, `Qwen3` in Table 10, and `RoBERTa` in Table 11.

C Additional Citation Purpose Annotation Details

Here, we present statistics on the distribution of citation purpose within the manually-annotated dataset across the citing work's year bucket (Figure 9) and venue (Figure 10).

D Common Disagreements in Citation Purpose Annotations

Here we discuss common disagreements in Citation Purpose annotation. Figure 3 shows all disagreements from our annotations study. Below we outline all common disagreements.

Related Work & Analysis There is some confusion between *Related Work* and *Analysis*. Commonly, this is because a comparison to the citing work, which would constitute *Analysis*, is subtle. Consider the below example of one such disagreement, taken from Hardalov et al. (2020), where the in-text reference in question is shown in bold:

> *Knowing that the context retrieved from the noisy Wikipedia corpus is relevant for answering Exodus questions, suggests that we need a better multilingual science corpus, similar to Clark et al. (2018); **Pan et al. (2019)**; Bhakthavatsalam et al. (2020). We further need better multilingual knowledge selection and ranking (Banerjee et al., 2019).*

While the authors are indeed making a positive comparison to a pre-existing corpus, it is easily overlooked. We note that positive comparisons are somewhat more difficult to detect than negative ones.

Related Work & Substantiation There is also some confusion between *Related Work* and *Substantiation*. This is generally due to less obvious claims, making them difficult to identify (as suggested in Part III of the codebook in Appendix 1). In the below excerpt from Masud et al. (2024), the reference is used to substantiate one part of a larger claim:

> *Disparity in access to additional context (Ljubešić et al., 2022;İhtiyar et al., 2023) and annotation guidelines **(Ross et al., 2016)**, in some instances, reduce the bias*

Section	Keywords
Related Work	"related work," "background," "related research"
Method	"method," "approach," "notation," "technique," "algorithm," "architecture," "design," "solution," "model," "corpus," "data," "framework"
Experiments	"experiment," "evaluation," "test," "analysis," "vs.," "compare," "accuracy," "scores," "state-of-the-art," "baseline," "results," "performance"
Conclusion	"conclusion," "future work," "discussion," "limitation," "ethical," "ethics"
Appendix	"appendix," "appendices"

Table 8: Keywords used for matching section titles to canonical sections.

Class	Precision	Recall	F1-Score	Support
substantiation + basis	0.00	0.00	0.00	10
substantiation	0.50	0.39	0.44	44
basis	0.07	0.38	0.12	8
definition	0.05	0.67	0.10	6
analysis	0.44	0.29	0.35	24
use	0.00	0.00	0.00	19
related work	1.00	0.02	0.03	57
accuracy			0.19	168
macro avg	0.30	0.25	0.15	168
weighted avg	0.54	0.19	0.19	168

Table 9: `Llama3` classification report showing precision, recall, and F1-score per citation purpose category.

Class	Precision	Recall	F1-Score	Support
substantiation + basis	0.20	0.20	0.20	10
substantiation	0.67	0.50	0.57	44
basis	0.38	0.75	0.50	8
definition	0.80	0.67	0.73	6
analysis	0.57	0.33	0.42	24
use	0.79	0.79	0.79	19
related work	0.65	0.81	0.72	57
accuracy			0.61	168
macro avg	0.58	0.58	0.56	168
weighted avg	0.62	0.61	0.60	168

Table 10: `Qwen3` classification report showing precision, recall, and F1-score per citation purpose category.

> *and, in some cases, confirm the annotator's biases.*

We argue that this could be easily mistaken for an example, constituting a *Related Work* classification.

Substantiation & Substantiation + Basis Confusion between *Substantiation* and *Substantiation + Basis* is also relatively common. This is relatively intuitive, as it's relatively simple to identify the claim constituting a *Substantiation* classification. However, language suggesting that this claim is explicitly being *built on* is subtler. Take the below excerpt from Desai et al. (2020) as an example.

Class	Precision	Recall	F1-Score	Support
substantiation + basis	0.00	0.00	0.00	10
substantiation	0.67	0.70	0.69	44
basis	0.22	0.25	0.24	8
definition	0.60	0.50	0.55	6
analysis	0.40	0.33	0.36	24
use	0.86	1.00	0.93	19
related work	0.63	0.72	0.67	57
accuracy			0.62	168
macro avg	0.48	0.50	0.49	168
weighted avg	0.58	0.62	0.60	168

Table 11: `RoBERTa` classification report showing precision, recall, and F1-score per citation purpose category.

> *Domain adaptation techniques address these challenges by efficiently using large amounts of unlabeled target domain data, consequently outperforming the aforementioned supervised techniques (**Alam et al., 2018**;Li et al., 2017). Our work contributes to disaster-centric emotion detection in three ways by: (1) introducing a dataset large enough to train supervised classifiers; (2) exploring various forms of pre-training to instill strong inductive biases...*

The claim that "domain adaptation techniques outperform prior supervised techniques" is relatively straightforward to identify. However, the citing work also explicitly aims to build on the technique(s) proposed in *Alam et al., 2018* to make them more effective, constituting a *Substatiation + Basis* classification.

E Replicating Engagement Analysis Results on Annotated Data

Because the the analysis in Section 6 is conducted over automatic citation purpose *predictions*, we replicate results on our smaller dataset of citations manually-annotated with citation purpose (de-

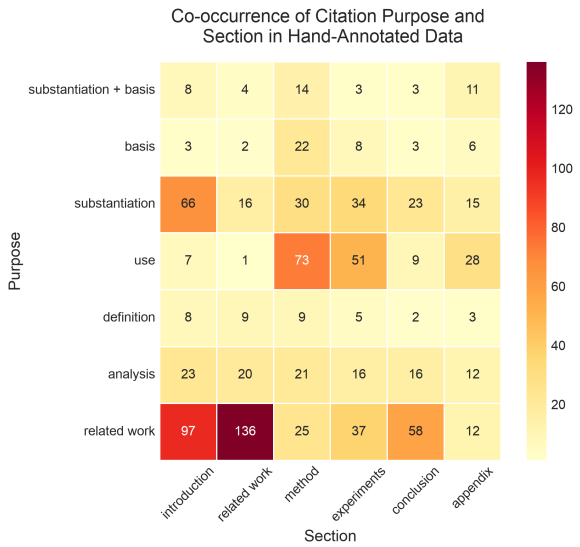

Figure 11: Co-occurrence of citation purpose and section in the annotated dataset reflects trends in the full dataset.

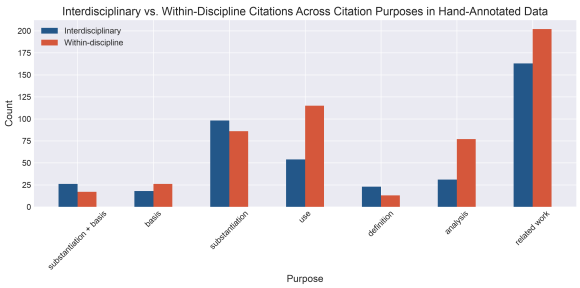

Figure 12: Interdisciplinary vs. Within-Discipline citations by citation purpose in hand-annotated dataset reflects trends in the full dataset.

scribed in Section 4.2). This helps to improve confidence in our findings.

To further support our claim that sections and citation purposes intuitively co-occur, we replicate the trends shown in Figure 5 on the hand-annotated purpose labels in Figure 11. Citations with a *Related Work* purpose label frequently appear in the *Introduction* and *Related Work* sections. *Substantiation*-purpose citations occur in the *Introduction* and *Use*-purpose citations appear in the *Method* and *Experiments* sections. Some of the weaker trends found in the full dataset, such as a greater frequency of *Analysis* citations in the *Introduction* and *Related Work* sections, are also found in the annotated data

We also include Figure 12 to show that our claim that the *Substantiation* and *Definition* citation purposes are more associated with interdisciplinary citations holds in the annotated dataset.

F Error Analysis

Here we outline potential failure points in our data collection and analysis framework.

F.1 Failure to Match In-text References to Corresponding Citations

Typically, when you click on an in-text reference in an ACL paper, you are linked to the corresponding entry in the References section. Grobid captures this information, extracting a navigation id which allows us to match in-text references to corresponding citations. However, sometimes this navigation id is missing, either due to parsing errors, or because a paper author did not properly format an in-text reference. Because of this, not all in-text references are properly extracted and matched to citations.

F.2 Failure to Match Citations to Field-of-Study

Sometimes citations in the *References Section* do not contain enough metadata to match them to confidently match them to a field of study due to parsing errors or errors in citations. If we are unable to map citations to a field of study using the Semantic Scholar API or string-matching approach, we do not consider them in our analysis.

F.3 Incorrect Section Assignments

The string matching approach used to match section titles to canonical sections (detailed in Appendix A.2) can lead to incorrect assignments. Substring matching on section titles can be misleading. For example, the keyword "approach" is used to map to *Method*. However, it is possible that an *Experiments* section title would contain the substring, such as in "Our Approach vs. Prior Work".

F.4 Incorrect Citation Purpose Assignments

Because we use an automatic classifier to predict Citation Purpose categories for our analysis in Section 6, it carries some uncertainty. We detailed the performance of the classifier in Section 4.3 to acknowledge this.

G Codebook

Our Citation Purpose Codebook is included in Appendix 1. We also include a pdf version along with our codebase and data.

Citation Purpose Codebook

The 7 categories below describe possible purposes for an out-of-discipline citation in an academic paper. Each category contains a description and guidelines for when the category is pertinent, an example, and an explanation of why that particular category is right for the example.

I. Use The method or tool in the citation is used directly in the citing work, with no or very trivial modifications.

Abstract: In this work, we carry out a data archaeology to infer books that are known to ChatGPT and GPT-4 using a name cloze membership inference query. We find that OpenAI models have memorized a wide collection of copyrighted materials, and that the degree of memorization is tied to the frequency with which passages of those books appear on the web. The ability of these models to memorize an unknown set of books complicates assessments of measurement validity for cultural analytics by contaminating test data; we show that models perform much better on memorized books than on nonmemorized books for downstream tasks. We argue that this supports a case for open models whose training data is known.
Section: experiments
Citation Context: Our second experiment predicts the amount of time that has passed in the same 250-word passage sampled above, using the conceptual framework of **Underwood (2018)**, explored in the context of GPT-4 in Underwood (2023). Unlike the year prediction task above, which can potentially draw on encyclopedic knowledge that GPT-4 has memorized, this task requires reasoning directly about the information in the passage, as each passage has its own duration. Memorized information about a complete book could inform the prediction about all passages within it (e.g., through awareness of the general time scales present in a work, or the amount of dialogue within it), in addition to having exact knowledge of the passage itself. As above, our core question asks: do we see a difference in performance between books that GPT-4 has memorized and those that it has not? To assess this, we again identify the top and bottom decile of books by their GPT-4 name cloze accuracy. We manually annotate the amount of time passing in a 250-word passage from that book using the criteria outlined by **Underwood (2018)**, including the examples provided to GPT-4 in Underwood (2023), blinding annotators to the identity of the book and which experimental condition it belongs to. We then pass those same passages through GPT-4 using the prompt in figure 4, adapted from Underwood (2023). We calculate accuracy as the Spearman ρ between the predicted times and true times and calculate a 95% confidence interval over that measure using the bootstrap.

The above citation refers to a framework that is used directly in the citing work. There is no explicit reference to modifying this framework, or building upon it, and so it is inferred that the framework is used as is, and thus categorized as 'Use' and not 'Basis'.

II. Definition The citation is used to define a topic or phrase (but NOT validate it - see 'Substantiation').

Abstract: This paper analyses two hitherto unstudied sites sharing state-backed disinformation, Reliable Recent News (rrn.world) and WarOnFakes (waronfakes.com), which publish content in Arabic, Chinese, English, French, German, and Spanish. We describe our content acquisition methodology and perform cross-site unsupervised topic clustering on the resulting multilingual dataset. We also perform linguistic and temporal analysis of the web page translations and topics over time, and investigate articles with false publication dates. We make publicly available this new dataset of 14,053 articles, annotated with each language version, and additional metadata such as links and images. The main contribution of this paper for the NLP community is in the novel dataset which enables studies of disinformation networks, and the training of NLP tools for disinformation detection.
Section: introduction
Citation Context: Propaganda is defined as content that intentionally influences opinion to advance its creators' goals **(Bolsover and Howard, 2017)**. Numerous propaganda datasets have previously been created, with both document-level (Rashkin et al., 2017;Barrón-Cedeño et al., 2019) and span-level (Da San Martino et al., 2019b) technique annotations, using articles collected from multiple disinformation sites. At article-level, classifiers using combinations of multiple linguistic representations based on style and readability outperform content representation (Barrón-Cedeño et al., 2019), whereas content-based transformer models such as BERT have seen use at span-level (Da San Martino et al., 2019a). Detectors are often evaluated on single datasets, prompting concerns on generalisation (Martino et al., 2020).

The above citation offers a definition for "propaganda" that is used in the citing work.

III. Substantiation The citation is used to verify or substantiate theoretical, empirical or methodological claims made in the citing work. Claims that describe the state of the literature (e.g. "There exist studies on this subject", "Multiple works have adopted this approach") are NOT valid claims for this category, and these citations might do better under 'Examples' or 'Related Work.' It does NOT matter if the claim is central to the premise of the citing work (such as claims made about contributions, or any claims in the abstract), or if the claim is local to the citation context. This category applies in any case where the citation is being used as proof of the claim. If classifying as 'Substantiation,' it is useful to identify the claim that is being supported.

Abstract: Prior work has revealed that positive words occur more frequently than negative words in human expressions, which is typically attributed to positivity bias, a tendency for people to report positive views of reality. But what about the language used in negative reviews? Consistent with prior work, we show that English negative reviews tend to contain more positive words than negative words, using a variety of datasets. We reconcile this observation with prior findings on the pragmatics of negation, and show that negations are commonly associated with positive words in negative reviews. Furthermore, in negative reviews, the majority of sentences with positive words express negative opinions based on sentiment classifiers, indicating some form of negation.
Section: introduction
Citation Context: A battery of studies have validated the Pollyanna hypothesis that positive words occur more frequently than negative words in human expressions, using corpora ranging from Google Books to Twitter (Dodds et al., 2015;Garcia et al., 2012;Boucher and Osgood, 1969; **Kloumann et al., 2012**). The typical interpretation is connected with the positivity bias, which broadly denotes 1) a tendency for people to report positive views of reality, 2) a tendency to hold positive expectations, views, and memories, and 3) a tendency to favor positive information in reasoning (Carr, 2011; Augustine et al., 2011; Hoorens, 2014). However, it remains an open question whether the Pollyanna hypothesis holds in negative reviews, where the communicative goal is to express negative opinions.

The citation offers proof of the claim "positive words occur more frequently than negative words in human expressions." While the fact of this evidence is mentioned explicitly in the citation context above, this explicit mention is not necessary to apply this category to a citation.

IV. Basis The method, idea, or tool used in the citation is used as a basis for the method, idea, or tool in the citing work. It must be explicitly stated that this is being built upon, or it must be very evident from the abstract provided that this is an idea that is central to the premise and is thus being explored further.

Abstract: Online misogyny is a pernicious social problem that risks making online platforms toxic and unwelcoming to women. We present a new hierarchical taxonomy for online misogyny, as well as an expert labelled dataset to enable automatic classification of misogynistic content. The dataset consists of 6,567 labels for Reddit posts and comments. As previous research has found untrained crowdsourced annotators struggle with identifying misogyny, we hired and trained annotators and provided them with robust annotation guidelines. We report baseline classification performance on the binary classification task, achieving accuracy of 0.93 and F1 of 0.43. The codebook and datasets are made freely available for future researchers.
Section: method
Citation Context: This taxonomy draws on the typologies of abuse presented by Waseem et al. (2017) and **Vidgen et al. (2019)** as well as theoretical work in online misogyny research (Filipovic, 2007;Mantilla, 2013;Jane, 2016;Ging, 2017;Anzovino et al., 2018;Ging and Siapera, 2019;Farrell et al., 2019). It was developed by reviewing existing literature on online misogyny and then iterating over small samples of the dataset. This deductive-inductive process allowed us to ensure that conceptually distinct varieties of abuse are separated and that different types of misogyny can be unpicked. This is important given that they can have very different impacts on victims, different causes, and reflect different outlooks and interests on the part of the speaker.

The citation is classified as 'Basis' because the authors state that their taxonomy (central to their work, although this factor is not necessary) "draws on" the typology presented in the citation.

V. Substantiation + Basis The citation is used as proof of a claim made in the citing work (see category III above), and this claim is either directly built upon in the citing work, or used as the basis for a specific idea (see category IV above).

Abstract: Multiple studies have demonstrated that behavior on internet-based social media platforms can be indicative of an individual's mental health status. The widespread availability of such data has spurred interest in mental health research from a computational lens. While previous research has raised concerns about possible biases in models produced from this data, no study has quantified how these biases actually manifest themselves with respect to different demographic groups, such as gender and racial/ethnic groups. Here, we analyze the fairness of depression classifiers trained on Twitter data with respect to gender and racial demographic groups. We find that model performance systematically differs for underrepresented groups and that these discrepancies cannot be fully explained by trivial data representation issues. Our study concludes with recommendations on how to avoid these biases in future research.
Section: method
Citation Context: Tokenization. Raw text within in Tweets was tokenized using a modified version of the Twokenizer (O'Connor et al., 2010). English contractions were expanded, while specific retweet tokens, username mentions, URLs, and numeric values were replaced by generic tokens. As pronoun usage tends to differ in individuals living with depression **(Vedula and Parthasarathy, 2017)**, we removed any English pronouns from our stop word set (English Stop Words from nltk.org). Case was standardized across all tokens, with a single flag included if an entire post was made in uppercase letters.

The citation offers proof for the claim "pronoun usage tends to differ in individuals living with depression," and this claim is used to justify their removal of English pronouns in their approach.

VI. Analysis The method or idea in the citation is being analyzed in some way by presenting comparisons or critiques. These analyses must have a direct link with the citing work, as evidenced in the citation context or the abstract. For example, the citation's idea or method may be compared to one in the citing work, or may be criticized and its limitations used to justify an approach taken in the citing work. There must be no explicit mention of building on the citation's idea in the citing work.

Abstract: Natural language inference (NLI) is the task of determining whether a piece of text is entailed, contradicted by or unrelated to another piece of text. In this paper, we investigate how to tease systematic inferences (i.e., items for which people agree on the NLI label) apart from disagreement items (i.e., items which lead to different annotations), which most prior work has overlooked. To distinguish systematic inferences from disagreement items, we propose Artificial Annotators (AAs) to simulate the uncertainty in the annotation process by capturing the modes in annotations. Results on the CommitmentBank, a corpus of naturally occurring discourses in English, confirm that our approach performs statistically significantly better than all baselines. We further show that AAs learn linguistic patterns and context-dependent reasoning.
Section: method
Citation Context: If we view the AAs as a committee of three members, our architecture is reminiscent of the Query by Committee (QBC) **(Seung et al., 1992)**, an effective approach for active learning paradigm. The essence of QBC is to select unlabeled data for labeling on which disagreement among committee members (i.e., learners pre-trained on the same labeled data) occurs. The selected data will be labeled by an oracle (e.g., domain experts) and then used to further train the learners. Likewise, in our approach, each AA votes for an item independently. However, the purpose is to detect disagreements instead of using disagreements as a measure to select items for further annotations. Moreover, in our AAs, the three members are trained on three disjoint annotation partitions for each item (see Section 3.2).

The citation refers to an architecture that is similar to the one proposed in the citing work. It is clear from the citation context that the authors of the citing work are contrasting their own architecture with the one in the citation.

VII. Related Work The citation refers to work that is related to the citing work or provides it as an example of a method, idea, or phenomenon discussed without further description. It does not fit any of the other categories.

Abstract: Methods and applications are inextricably linked in science, and in particular in the domain of text-as-data. In this paper, we examine one such text-as-data application, an established economic index that measures economic policy uncertainty from keyword occurrences in news. This index, which is shown to correlate with firm investment, employment, and excess market returns, has had substantive impact in both the private sector and academia. Yet, as we revisit and extend the original authors' annotations and text measurements we find interesting text-as-data methodological research questions: (1) Are annotator disagreements a reflection of ambiguity in language? (2) Do alternative text measurements correlate with one another and with measures of external predictive validity? We find for this application (1) some annotator disagreements of economic policy uncertainty can be attributed to ambiguity in language, and (2) switching measurements from keyword-matching to supervised machine learning classifiers results in low correlation, a concerning implication for the validity of the index.
Section: experiments
Citation Context: KeyExp. Although Baker et al. use human auditors to find policy keywords that minimize the false positive and false negative rates, they do not expand or optimize for economy or uncertainty keywords. Thus, we expand these keyword lists via GloVe word embeddings 9 (Pennington et al., 2014), and find the five nearest neighbors via cosine distance. 10 This is a simple keyword expansion technique. In future work, one could look to the literature on lexicon induction to improve creating lexicons that represent the semantic concepts of interest (Taboada et al., 2011;Pryzant et al., 2018;**Hamilton et al., 2016**;Rao and Ravichandran, 2009). Alternatively, one could also create a probabilistic classifier over pre-selected lexicons to soften the predictions, or use other uncertainty lexicons or even automatic uncertainty cue detectors.

As inferred from the given context, the citation describes work related to creating lexicons that represent the semantic concepts of interest, which is identifiable as related to the citing work. This is classified as 'Related Work,' and not 'Examples,' because there are multiple ideas being discussed, and it is not clear which, if any, the citation is supposed to be an example of. Additionally, the ideas being referenced are potential future directions, implying a certain level of novelty; it is therefore unlikely that the citation contains a concrete example of the specified idea.

Codebook 1: Codebook for Citation Purpose annotations.

H Prompts

The prompt used for Citation Purpose prediction are included in Appendix H.1 and H.2.

System Prompt

You are an expert in academic citation analysis. Your task is to classify the purpose of a citation in a scientific paper.

User Prompt

Given the following citation context and citation text, choose the most appropriate category from the options below. Each option also has two examples to help you understand the category. Please read the category descriptions and examples carefully before making your choice.

Citation context: <citation context>

Citation text: <citation text>

Candidate categories: <categories>

Respond with ONLY the category name in a JSON object, exactly as written above. Do not include any explanation.

Here is the expected output format: {“category”: “category name”}

Prompt H.1: Predicting Citation Purpose.

substantiation + basis

1. The citation implicitly or explicitly provides evidence for a theoretical, empirical, or methodological claim.
2. The above is true AND the claim does NOT describe the state of the literature (e.g. “There exist studies on this subject”, “Multiple works have adopted this approach”).
3. The above are true AND the claim is NOT being used as the basis for the citing work’s idea or method.
4. The above are true AND the citation includes a specific method or idea that is being used, built upon, improved, or otherwise modified.
5. The above are true AND the citation is NOT used directly.

substantiation

1. The citation implicitly or explicitly provides evidence for a theoretical, empirical, or methodological claim.
2. The above is true AND the claim does NOT describe the state of the literature (e.g. “There exist studies on this subject”, “Multiple works have adopted this approach”).
3. The above are true AND the claim is NOT being used as the basis for the citing work’s idea or method.

basis

1. The citation refers to an idea being used as a basis by the citing work.
2. The citation refers to a method that is explicitly being built upon, improved, or modified by the citing work. It is NOT used directly.
3. There is no explicit mention but it is very obvious from the abstract that this is integral to the citing work’s idea and is thus being built upon.
4. One of the above is true AND there is NO theoretical or empirical claim associated with the idea or method that motivates its use.

definition

1. The citation helps define or explain a topic or phrase.
2. The above is true AND there is no justification, validation, or proof otherwise provided.

analysis

1. The context offers a critique of the citation’s work.
2. The citation’s work is explicitly compared to the citing work.
3. The citation’s work is being used as a baseline to compare with the citing work’s approach.
4. One of the above is true, with NO explicit mention of building upon, using, or modifying the citation’s work.

use

1. The citation refers to a specific concrete method, not an idea, being used by the citing work.
2. The above is true AND there is no explicit mention of building upon, improving, or modifying the citation’s method in a non-trivial way.

related work

1. The citation is related to the article.
2. The above is true AND the citation does not fall under any of the other categories.

Prompt H.2: Category descriptions included in prompt.