Citation Context Classification

Thi-Minh-Thu Vu
Introduction to Cognitive Intelligence - INT3421 55
22028116@vnu.edu.vn

***Abstract*—Citations have long been essential indicators of influence and progress across various scientific fields, with citation counts frequently serving as a metric for assessing the impact of research work. Furthermore, the similarity between abstracts has proven to be a valuable tool in predicting the relevance of a document. Despite numerous efforts to automate the classification of citation importance, significant challenges remain. This paper offers an overview of the 2021 3C task, part of the Scientific Document Processing (SDP 2021) workshop. The task was split into two primary components: classifying the purpose of citations (Task A) and assessing their level of influence (Task B). In our study, we employed a diverse range of methods to categorize citation contexts, using approaches such as rule-based techniques, traditional machine learning algorithms, and advanced deep learning models. Results indicated that the fine-tuned SciBERT model outperformed all other systems within the deep learning category. For traditional machine learning, the fine-tuned RoBERTa and RandomForest models also delivered solid results, though they did not surpass the benchmark set by SciBERT.**

***Index Terms*—citation, multi-document summarization, web crawl, extractive summarization, query-driven**

I. Introduction

Over the past few years, the volume of scientific publications and research data available online has grown substantially, reflecting advancements across various fields. However, authors use citations for different purposes, and this diversity affects how future scholars interpret them. As Garfield (1972) points out [4], authors cite research for a variety of reasons, meaning that citations in a research paper should not be treated as identical.

Evaluating research based on citation context can help determine the impact of a study on the research community, as well as assess the reliability and credibility of that research. Research evaluation should consider not only the quantity of citations but also their context and purpose. This approach helps readers better understand the context and meaning of citations, making it easier to analyze information sourced from various references. Studies in this area can enhance our understanding of scientific development and contribute to building a common knowledge base.

Over time, citation classification methods have shifted from manual, human-based evaluations to automated techniques that utilize progress in machine learning and natural language processing. These classifications have become increasingly complex, considering not just the surface role of citations but also their context, purpose, and scientific influence, identifying and analyzing the relationships between citations and their contexts. This evolution has optimized the citation classification process, facilitating faster and more efficient scientific document analysis.

The rise of aggregation platforms such as CORE has significantly expanded open access to academic literature. The accessibility of complete research documents enhances the scalability of bibliometric studies by providing clearer insights into citation contexts. Collaborative efforts during SDP 2021 concentrated on categorizing citation contexts in research papers according to their purpose and impact.

To achieve accurate classification of citation functions in scientific studies, various methods and classification schemes have been developed based on manual evaluation and analysis of research papers. Significant contributions to the field include a classification of citations by their substantive nature and methodology, as well as an emphasis on the purpose of citations as either supportive or opposing [2], [9]. Another analysis focused on the impact level of citations, ranging from ordinary to highly influential [13]. Teufel et al. (2006) [14] developed a scheme with 12 functional layers for automatic classification. Subsequent works introduced innovative approaches for understanding citation contexts, significantly enhancing the classification methods used in research [1]. Building on this foundation, further refinements of these techniques incorporated more nuanced criteria for assessing citation functions [8]. Following this, additional frameworks were developed that leveraged deep learning to improve the accuracy and efficiency of citation classification [3]. Finally, an emphasis on the importance of contextual analysis demonstrated the broad applicability of these methods across various academic disciplines [11]

Despite significant efforts and over a decade of research, comparing and contrasting current methods in this area remains challenging due to the existence of multiple classification systems and the lack of a common standard. To address this issue, the 3C sharing task was introduced to analyze citation sentences. This task is divided into two parts: Subtask A focuses on classifying citation sentences by their purpose or function, while Subtask B aims to evaluate the influence or impact of citations. Both tasks make use of a subset of the Academic Citation Classification dataset, as detailed in previous studies [11], [12].

II. Related Work

Over the past few decades, there have been significant efforts to automate the process of citation function classification in scientific papers. This is necessary due to the growing number of academic publications, making manual citation tracking and analysis increasingly impractical. Most contemporary automated systems depend on taxonomies and artificial intelligence techniques to determine citation functions and evaluate their significance in relation to the citation's context.

Multiple classification frameworks for citation functions have been created to examine the intent behind citations. These schemes often use predefined categories to determine the citation's role, ranging from providing background to comparing studies or critiquing previous results. These classification systems vary in complexity, with some containing a few classes (2–3), others having a moderate number (5–10), and some featuring over 12 classes.

- *Teufel et al. (2006)* [14] developed a 12-class citation function taxonomy, which is considered one of the more complex systems, especially focused on the citation's role within the specific context of the scientific paper.
- Building on this approach, significant contributions to the development of citation function classification schemes were made in 2013 [1]. Further advancements in these ideas occurred in 2018, when automated systems were created that incorporated specific categories, as noted in the literature [8] further advanced these ideas by creating automated systems that included specific categories like "supportive," "comparison," or "critique."

Other taxonomies have included the polarity aspect of citations, focusing on whether they are positive (supportive) or negative (oppositional). In-depth examinations of citation polarity were carried out in 2013 [1] and further explored in 2015 [5], providing valuable insights into the positive and negative interactions among academic works. Understanding citation polarity is critical for grasping the relationships and arguments between scientific studies within the academic community.

In addition to function classification, another significant aspect of citation classification is the influence that citations exert. Influence classification methods generally employ a binary framework, categorizing citations as either influential or non-influential. This classification is useful for assessing the impact a specific study has on the wider scientific community. Notably, a pivotal study conducted in 2015 focused on developing influence classification methods [16]. Similarly, other research contributions from the same year also played a significant role in this area [15]. Both studies employ a binary approach to determine whether a citation significantly impacts other research, forming a basis for evaluating citation importance.

Early methods for classifying citations relied on supervised machine learning models that integrated various features. These models utilized attributes extracted from the citation context, encompassing elements such as vocabulary, grammar, and the interrelationships among different studies. Notably, models like Random Forest (RF) were implemented in 2018 [8] and in another significant study conducted in 2017 [10]. Conversely, Support Vector Machine (SVM) techniques were utilized in 2017 in separate research efforts [6] and [7]. Nevertheless, both methodologies required researchers to manually identify contextual features before training the models, which rendered the process labor-intensive and impractical for larger datasets.

The emergence of larger datasets, such as SciCite and Academic Citation Classification, has facilitated significant advancements in citation classification. Notably, the SciCite dataset was introduced in 2019 [3], while contributions to the Academic Citation Classification dataset were made in 2020 [11] and in 2019 [12]. These datasets have enabled the development of more sophisticated models. Among these advancements, transformer-based architectures, particularly SciBERT, which was developed based on BERT and specifically trained on scientific texts, have been employed to enhance the accuracy and efficiency of automatic citation classification. SciBERT has notably improved performance in the analysis and classification of citation functions.

III. Materials and Methods

Recent efforts in the field of scholarly document processing have increasingly emphasized the importance of classifying citation contexts within research publications. This classification aims to analyze the influence and purpose of citations, providing valuable insights into the dynamics of academic discourse and the relationships among various research works.

A. Dataset

The ACT dataset comprises 3,000 training examples that showcase citation contexts from a wide array of academic fields, making it a valuable resource for researchers seeking to understand citation dynamics. Each citation within this dataset is meticulously categorized according to its purpose or influence, allowing for a nuanced analysis of how different types of citations contribute to scholarly discourse.

However, it is important to note that the dataset exhibits a significant class imbalance. A disproportionate number of citations are classified as either "BACKGROUND" or "INCIDENTAL", which skews the representation of citation types. This imbalance can complicate the classification task, as models may become biased towards the more prevalent categories, leading to a decrease in their ability to accurately classify less common citation types.

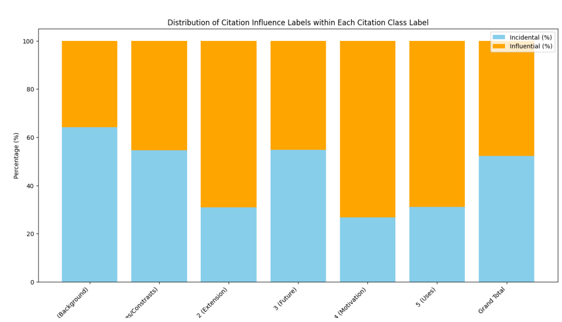

Fig. 1. Distribution of Citation Influence within Each Citation Class

An example from the training dataset is shown in Table 1.

TABLE I
A SAMPLE ENTRY FROM THE TRAINING DATASET.

unique_id	core_id	citing_title	citing_author	cited_title	cited_author	citation_context	citation_class_label	citation_influence_label
1998	81605842	Everolimus improves behavioral deficits in a patient with autism associated with tuberous sclerosis: a case report	Ryouhei Ishii	Learning disability and epilepsy in an epidemiological sample of individuals with tuberous sclerosis complex	Joinson	West syndrome (infantile spasms) is the commonest epileptic disorder, which is associated with more intellectual disability and a less favorable neurological outcome (#AUTHOR_TAG et al, 2003)	4	1

TABLE II
THE PROPORTIONAL DISTRIBUTION OF THE LABELS.

Count of unique_id	Column Labels			Incidental Count	Influential Count	
Row Labels	**0 (Incidental %)**	**1 (Influential %)**	**Grand Total**	**0**	**1**	**Grand Total**
0 (Background)	64.20%	35.80%	100.00%	1058	590	1648
1 (Compares/Constrasts)	54.62%	45.38%	100.00%	201	167	368
2 (Extension)	30.99%	69.01%	100.00%	53	118	171
3 (Future)	54.84%	45.16%	100.00%	34	28	62
4 (Motivation)	26.81%	73.19%	100.00%	74	202	276
5 (Uses)	31.16%	68.84%	100.00%	148	327	475
Grand Total	**52.27%**	**47.73%**	**100.00%**	**1568**	**1432**	**3000**

B. Overview

The 3C shared task presents a citation classification challenge, divided into two distinct subtasks: The frequency of these labels is also presented in Table II and The distribution of the data fields in the dataset is as follows Fig. 1:

1) Subtask A: The goal of this subtask is to determine the function of a citation. Participants will categorize citations into one of six specified classifications: BACKGROUND, USES, COMPARES_CONTRASTS, MOTIVATION, EXTENSION, and FUTURE.

2) Subtask B: This subtask focuses on assessing the relevance of a citation. It involves a dual-classification approach, in which citations are required to be categorized as either INCIDENTAL or INFLUENTIAL.

C. The Proposed Model

In task A, which involves the task of classifying citation purposes, we used the SciBERT model to classify citations. We focused on learning from citation contexts, specifically the sentences containing citations, without utilizing other information (such as titles or other sections of the paper). This approach helped me achieve the highest score.

We fine-tuned the SciBERT model, both the cased and uncased versions, combined with a linear layer for prediction. To address the class imbalance issue (where some classes had significantly fewer data than others), we applied a weighted loss function and leveraged extended contextual information, enhancing the classification performance for the underrepresented classes.

Instead of merely using a simple linear neural net layer after the SciBERT model, we added an LSTM layer. As we know, SciBERT is highly effective at capturing relationships and context within a text sequence, but it is often paired with a linear layer to produce the final classification. In contrast, LSTM is a recurrent neural network known for its ability to retain long-term information and better handle sequential data, making it particularly effective at understanding context and word order in text sequences.

By combining SciBERT and LSTM, the model can extract deep features from the text, understanding the relationships between words and the context within sentences, and better capturing long-term dependencies and word order within the text sequence. This improves the model's ability to handle long and complex text sequences while retaining important information from previous context. As a result, the accuracy of classifying citation contexts is enhanced, particularly in cases that require understanding the connections between different parts of the text.

D. Class-weighted training

With the class-weighted training approach, we used a deep learning model like SciBERT and incorporated class weights into the loss function to handle data imbalance. The code demonstrates how class-weighted training is applied, ensuring that underrepresented classes still have a strong influence during the training process. Without class weights, the model could become biased towards classes with more samples, as these are easier to achieve higher accuracy.

In implementing the weighted loss calculation, we made use of the Cross-Entropy Loss method, which included class weights. This function allows the model to learn how to classify underrepresented classes more accurately by penalizing prediction errors more heavily for these classes.

The formula for Cross-Entropy with class weights is as follows:

$$L(y, \hat{y}) = -\sum_{i=1}^{C} w_i \cdot y_i \cdot \log(\hat{y}_i) \quad (1)$$

where C represents the total number of classes, w_i signifies the weight associated with class i, y_i indicates the actual label, and $\hat{y}_i$ denotes the probability predicted for that class.

For model optimization, we used the Adam optimizer. This technique combines Momentum and RMSProp, adjusting the model's weights based on the gradient of the loss function with a learning rate (`LEARNING_RATE`). Adam helps the model converge faster and more efficiently, especially in classification tasks with imbalanced data, ensuring that the model does not become biased towards classes with more samples.

E. Evaluating the Model Using F1 Score

In the evaluation of classification models, particularly those dealing with imbalanced datasets, the F1 score stands out as a crucial performance metric. This score provides a balance between precision and recall, making it especially useful for assessing models where one class may be significantly under-represented. The F1 score is computed using the following formula:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2)$$

- **Precision** is the ratio of correctly predicted positive observations to the total predicted positives.
- **Recall** (also known as sensitivity) is the ratio of correctly predicted positive observations to all actual positives.

By combining precision and recall in a harmonic mean, the F1 score ensures that both metrics are accounted for equally. The score ranges from 0 to 1, where a value closer to 1 represents better model performance. This is particularly beneficial in scenarios where the dataset is skewed toward one class, preventing the model from being biased toward the majority class. This dual focus empowers researchers and practitioners to make more informed adjustments to their models, leading to a more thorough and reliable assessment of their classification performance.

IV. Experiments and Results

A. Linear Layer Fine-Tuning: BERT, RoBERTa, and SciBERT (Cased & Uncased)

Here are the results of Experiment 1. Please refer to Table III, noting that the name of the model is based on the model type, followed by the parameters: `TRAIN_BATCH_SIZE`, `LEARNING_RATE`, and `DROP_OUT`.

Refer to the results of the experiment in Table III and in Table IV.

The model `run1_bert-base-uncased_4_0.00001_0` achieves the highest Macro F1 Score, with the highest Macro F1 Score, demonstrating an improved equilibrium between precision and recall in comparison to the other models. Nevertheless, it does not hold the highest accuracy among them.

TABLE III
Model Performance Metrics - Task A

Model	Macro F1 Score	Accuracy
bert-base-uncased_4_0.00001_0.1	0.417995679	59.8
bert-base-uncased_4_0.00001_0	0.470888124	58.4
bert-base-uncased_8_0.00001_0.1	0.434368536	63.2
bert-base-uncased_8_0.00001_0	0.416701613	63.6
roberta-base_4_0.00001_0.1	0.426885187	63.0
roberta-base_4_0.00001_0	0.404356568	59.8
roberta-base_8_0.00001_0.1	0.418994303	52.8
roberta-base_8_0.00001_0	0.414432233	58.8
scibert_scivocab_cased_4_0.00001_0.1	0.422761151	58.6
scibert_scivocab_cased_4_0.00001_0	0.407681996	57.0
scibert_scivocab_uncased_4_0.00001_0.1	0.423137228	57.6
scibert_scivocab_uncased_4_0.00001_0	0.436147794	61.8
scibert_scivocab_cased_8_0.00001_0.1	0.41655102	61.8
scibert_scivocab_cased_8_0.00001_0	0.43836798	61.4
scibert_scivocab_uncased_8_0.00001_0.1	0.431022956	58.8
scibert_scivocab_uncased_8_0.00001_0	0.423979694	58.2

The model `run1_roberta-base_8_0.00001_0.1` has the lowest Macro F1 Score and also the lowest accuracy, suggesting it has poor classification capability and may require adjustments or improvements.

TABLE IV
Model Performance Metrics - Task B

Model	Macro F1 Score	Accuracy
bert-base-uncased_4_0.00001_0.1	0.655977983	65.6
bert-base-uncased_4_0.00001_0	0.663655583	66.4
bert-base-uncased_8_0.00001_0.1	0.434368536	63.2
bert-base-uncased_8_0.00001_0	0.416701613	63.6
roberta-base_4_0.00001_0.1	0.663994624	66.4
roberta-base_4_0.00001_0	0.655950457	65.6
roberta-base_8_0.00001_0.1	0.665774063	66.6
roberta-base_8_0.00001_0	0.667914986	66.8
scibert_scivocab_cased_4_0.00001_0.1	0.655977983	65.6
scibert_scivocab_cased_4_0.00001_0	0.651726954	65.2
scibert_scivocab_uncased_4_0.00001_0.1	0.423137228	57.6
scibert_scivocab_uncased_4_0.00001_0	0.655977983	65.6
scibert_scivocab_cased_8_0.00001_0.1	0.655647383	65.6
scibert_scivocab_cased_8_0.00001_0	0.669300239	67
scibert_scivocab_uncased_8_0.00001_0.1	0.677812856	68.2
scibert_scivocab_uncased_8_0.00001_0	0.676	67.6

The model `run1_scibert_scivocab_uncased_8_0.00001_0.1` demonstrates the best performance in the table, achieving the highest Macro F1 Score and Accuracy. This indicates that the model has excellent classification capabilities, particularly in tasks requiring high precision. SciBERT appears to perform effectively in this domain, likely due to its suitable architecture and training data. Conversely, the model's ability to handle nuanced citation contexts further enhances its reliability, making it a strong candidate for various citation classification challenges. Future work could explore fine-tuning parameters or expanding the dataset to assess potential improvements in performance even further.

B. Task 1 with an Unweighted Loss

This experiment focuses specifically on task 1, aiming to compare the results obtained from implementing both weighted and unweighted loss functions. The primary objective is to assess how the use of an unweighted loss function affects the overall performance of the model. Refer to the results of the experiment in Table V.

TABLE V
MODEL PERFORMANCE METRICS FOR SCIBERT (UNCASED)

Model	Macro F1 Score	Accuracy
scibert_scivocab_uncased_4_0.00001_0.1	0.420859867	60.4
scibert_scivocab_uncased_4_0.00001_0	0.426587673	64
scibert_scivocab_uncased_8_0.00001_0.1	0.426950434	62.2
scibert_scivocab_uncased_8_0.00001_0	0.422635177	61.4

The model `scibert_scivocab_uncased_4_0.00001_0` has the highest accuracy among the listed models.

C. SciBERT with LSTM Classification

The goal of this experiment is to explore the potential for improving classification performance by combining SciBERT and LSTM, thereby leveraging the strengths of both models in processing and analyzing text data.

Examine the experimental results presented in Table VI as well as Table VII.

TABLE VI
PERFORMANCE METRICS FOR SCIBERT MODELS - TASK A

Model	Macro F1 Score	Accuracy
scibert_scivocab_uncased_4_0.00001_0.1	0.392063443	57
scibert_scivocab_uncased_4_0.00001_0	0.415941775	62
scibert_scivocab_uncased_8_0.00001_0.1	0.405577486	59
scibert_scivocab_uncased_8_0.00001_0	0.40932039	62

TABLE VII
PERFORMANCE METRICS FOR SCIBERT MODELS - TASK B

Model	Macro F1 Score	Accuracy
scibert_scivocab_uncased_4_0.00001_0.1	0.659863946	66
scibert_scivocab_uncased_4_0.00001_0	0.671242543	67.2
scibert_scivocab_uncased_8_0.00001_0.1	0.676767677	68
scibert_scivocab_uncased_8_0.00001_0	0.671110683	67.2

D. Using Citing Title with Citation Context for Fine-Tuning SciBERT

Here, we concatenate the citing title along with the citation context and use it with an architecture similar to that of the first experiment (SciBERT with a linear layer). Check the results from the experiment outlined in Table VIII as well as Table IX.

TABLE VIII
PERFORMANCE METRICS FOR SCIBERT MODELS - TASK A

Model	Macro F1 Score	Accuracy
scibert_scivocab_uncased_4_0.00001_0.1	0.437246422	62.4
scibert_scivocab_uncased_4_0.00001_0	0.425588666	63.0

TABLE IX
PERFORMANCE METRICS FOR SCIBERT MODELS - TASK B

Model	Macro F1 Score	Accuracy
scibert_scivocab_uncased_4_0.00001_0.1	0.675820215	67.8
scibert_scivocab_uncased_4_0.00001_0	0.665023208	66.6

E. Using Random Forest for Classification

We endeavored to employ the random forest technique to classify the embeddings generated from SciBERT. The model's designation reflects two key hyperparameters: the upper limit on the depth of the trees and the total count of trees within the ensemble. This specification is crucial, as it directly influences the model's performance and its ability to generalize across various datasets. Refer to the results of the experiment in Table X and Table XI.

TABLE X
PERFORMANCE METRICS FOR RANDOM FOREST MODELS - TASK A

Model	Macro F1 Score	Accuracy
random_forest_try35_500	0.259630272	0.606
random_forest_try35_1000	0.269477817	0.614
random_forest_try35_1700	0.271943621	0.616
random_forest_try40_1000	0.267172639	0.612

TABLE XI
PERFORMANCE METRICS FOR RANDOM FOREST MODELS - TASK B

Model	Macro F1 Score	Accuracy
random_forest_try25_1200	0.655911913	0.656
random_forest_try30_1200	0.653965397	0.654
random_forest_try35_1000	0.651726954	0.652
random_forest_try40_1200	0.657932955	0.658

Through five experiments, we derived the highest and lowest values of the Macro F1 Score, as shown in Table XII.

TABLE XII
RESULT MACRO F1 SCORE FOR TASK A AND TASK B

Task	Max (Macro F1 Score)	Min (Macro F1 Score)
Task A	0.470888124	0.392063443
Task B	0.677812856	0.434368536

V. CONCLUSION

In conclusion, this study provides an in-depth examination of the 2nd 3C citation context classification task, which was a significant part of the SDP workshop at the NAACL conference. By establishing two distinct subtasks for categorizing citations based on their purposes and influences, we have expanded upon the groundwork established in the previous iteration of this challenge. This time, we observed a notable enhancement in private macro F-score values, indicating advancements in analytical methodologies driven by the integration of advanced deep learning techniques, a robust text dataset, and diverse external information sources.

However, the overall performance metrics highlight ongoing challenges, particularly stemming from the complexities associated with the dataset's multi-domain nature and the presence of multiple citations within individual citation contexts. These factors emphasize the necessity for further research and innovative approaches to improve classification accuracy. Future efforts should concentrate on tackling these intricate issues to develop more effective systems capable of accurately interpreting citation contexts across various academic fields. The insights gained from this task not only contribute to our understanding of citation dynamics but also lay the groundwork for further enhancements in citation analysis methodologies.

ACKNOWLEDGMENT

We would like to thank Dr. Nguyen The Hoang Anh for providing valuable knowledge in the field of *Neuroscience*. We also thank all members of the Robotics Laboratory, UET, and VNU for their continuous support and encouragement.

REFERENCES

[1] Amjad Abu-Jbara, Jefferson Ezra, and Dragomir Radev. 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, 2013.

[2] Daryl E Chubin and Soumyo D Moitra. Content analysis of references: Adjunct or alternative to citation counting? *Social studies of science*, 5(4):423–441, 1975.

[3] Arman Cohan, Waleed Ammar, Madeleine Van Zuylen, and Field Cady. Structural scaffolds for citation intent classification in scientific publications. *arXiv preprint arXiv:1904.01608*, 2019.

[4] Eugene Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178(4060):471–479, 1972.

[5] Myriam Hernández-Alvarez and José M Gómez. Citation impact categorization: for scientific literature. In *2015 IEEE 18th International Conference on Computational Science and Engineering*, pages 307–313. IEEE, 2015.

[6] Myriam Hernandez-Alvarez, José M Gomez Soriano, and Patricio Martínez-Barco. Citation function, polarity and influence classification. *Natural Language Engineering*, 23(4):561–588, 2017.

[7] Rahul Jha, Amjad-Abu Jbara, Vahed Qazvinian, and Dragomir R Radev. Nlp-driven citation analysis for scientometrics. *Natural Language Engineering*, 23(1):93–130, 2017.

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

[9] Michael J Moravcsik and Poovanalingam Murugesan. Some results on the function and quality of citations. *Social studies of science*, 5(1):86–92, 1975.

[10] David Pride and Petr Knoth. Incidental or influential?–a decade of using text-mining for citation function classification. 2017.

[11] David Pride and Petr Knoth. An authoritative approach to citation classification. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, pages 337–340, 2020.

[12] David Pride, Petr Knoth, and Jozef Harag. Act: An annotation platform for citation typing at scale. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 329–330. IEEE, 2019.

[13] Ina Spiegel-Rosing. Science studies: Bibliometric and content analysis. *Social studies of science*, 7(1):97–113, 1977.

[14] Simone Teufel, Advaith Siddharthan, and Dan Tidhar. Automatic classification of citation function. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 103–110, 2006.

[15] Marco Valenzuela, Vu Ha, and Oren Etzioni. Identifying meaningful citations. In *Workshops at the twenty-ninth AAAI conference on artificial intelligence*, 2015.

[16] Xiaodan Zhu, Peter Turney, Daniel Lemire, and André Vellino. Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*, 66(2):408–427, 2015.