Predicting the presence of inline citations in academic text using binary classification

Abstract

Properly citing sources is a crucial component of any good-quality academic paper. The goal of this study was to determine what kind of accuracy we could reach in predicting whether or not a sentence should contain an inline citation using a simple binary classification model. To that end, we fine-tuned SciBERT on both an imbalanced and a balanced dataset containing sentences with and without inline citations. We achieved an overall accuracy of over 0.92, suggesting that language patterns alone could be used to predict where inline citations should appear in academic text.

1 Introduction

Providing accurate, relevant citations is an essential part of academic writing. Not only do citations allow authors to better contextualize the results of their paper, but they also lend credibility and authority to the claims made in the article. Failing to give credit to existing research when credit is due, on the other hand, is taken to show a lack of academic integrity, and is strongly frowned upon by the academic community. Appropriately adding citations, however, is not trivial: even humans sometimes struggle to determine where inline citations should go, and what should or should not be cited. This is particularly true in the case of junior academics and students (Vardi, 2012) (Carson et al., 1992) (Pennycook, 1996). In the context of *automatic* text evaluation, determining where citations should go is even less straightforward. One way in which one could automatically determine whether a given paragraph requires (additional) inline citations is through automatic plagiarism detection systems. However, processing a document to determine whether some sections of it have been plagiarized can require a considerable amount of time, particularly if the document exceeds a certain length. Building a plagiarism checker is also complicated, as the process requires scanning the full web for documents, and possibly obtaining access to research articles that might lay behind a paywall. Finally, results might not always be accurate (Kohl Kerstin, 2012), as the checker might fail in finding similarities between concepts simply because sentences that are identical in meaning have been expressed through a different formulation. Because of these downsides, we were interested in exploring how much mileage we could get out of a simple binary classification experiment trying to predict whether a given sentence should or should not include an inline citation. In particular, we reasoned that it should be possible to predict at least to some extent whether a sentence should contain an inline citation just by looking at the presence vs. absence of specific lexical cues. For example, verbs such as ”claimed”, nouns such as ”authors” and phrases such ”as seen in” tend to appear together or in the vicinity of inline citations. The same holds true of some capitalized nouns (e.g. ”Attention”, ”Minimalism”).

Developing shallow automated techniques that can detect whether or not a sentence should contain an inline citation has several practical applications. A shallow inline-citation predictor can be used to (i) help academics identify forgotten inline citations, i.e. citations that the author meant to add at the review stage but ultimately forgot to include, (ii) guide junior researchers in the paper-writing process, flagging concepts or ideas that might require attribution, (iii) improving the coverage of automatic essay analyzers, and (iv) in the context of natural language generation, decreasing the chances of committing plagiarism by flagging passages that might require a citation.

References play an essential role in academia and as such have been the target of several NLP

studies. In the last years, we have seen an increased tendency towards using references as a way to build knowledge graphs (Viswanathan et al., 2021) and speed up the search for relevant research articles. There is also a tendency towards using references to aid automated text summarization (Yasunaga et al., 2019). To our knowledge, however, this is the first study aiming to predict the need for citation in text without going through the need of processing huge amounts of research articles and performing plagiarism checks.

2 Preparing the Data

To determine what types of inline citation styles are used in different disciplines, we randomly selected two articles for each of the following 18 research fields: Medicine, Biology, Chemistry, Engineering, Computer Science, Physics, Math, Psychology, Economics, Political Science, Business, Geology, Sociology, Geography, Environmental Science, Art, History, Philosophy. After analyzing these 36 articles we concluded that most of the articles adopted the IEEE, APA or the Chicago reference styles.

We first created an initial dataset consisting of 2000 research articles; these were randomly selected from the ArXiv and PubMed datasets (Cohan et al., 2018) that are freely available on the Huggingface Datasets library (Lhoest et al., 2021) (`https://huggingface.co/datasets/scientific_papers`).

These 2000 articles were subsequently processed to discard articles with a citation pattern other than the IEEE, APA or Chicago reference styles. The pre-processing task of detecting inline citations was handled through a simple Python script. Using regular expressions, different kinds of citation styles were mapped to corresponding regex capture patterns. We started by writing regexes that would match the three citation styles that we identified as the most frequently used: IEEE, APA and Chicago. Later on, we also decided to include the alpha BibTeX style, as that appears to be quite frequently used in ArXiV papers. The Python script did the following: first, every given citation pattern was extracted from the article's plain text. Then, the style with the highest capture count was set as the article's default style. This means that even when the extraction process found inline citations that matched a style that was not the article's primary citation style, the script was still able to identify the primary style. Finally, the inline citations matching the primary style were substituted with an -ADD-CITATION- token; this step is important as it allowed us to generalize across different referencing styles. If for some reason no citation style was detected, the token replacement failed, and the article was discarded from further analysis.

We then created a second dataset by taking all the articles with IEEE, APA or Chicago as reference styles and by (i) breaking down the original text into sentences, and assigning each sentence to a separate entry, (ii) assigning different labels to entries containing inline citations and entries not containing inline citations, and (iii) removing the -ADD-CITATION- token throughout the dataset. This second dataset features 411'992 sentences (entries), of which 54'735 contain an inline citation (see Table 1) . The dataset is freely available at `https://github.com/elenaSage/InlineCitationSet`. This second dataset was the dataset we used for the classification experiments that we describe below.

Contains Citation	No Citation	Total
357257	54735	411992

Table 1: Composition of Inline Citation Database

3 Classification model

In recent years, BERT-based language models (Devlin et al., 2018) have achieved state-of-the-art performance in numerous NLP classification tasks. Due to their pre-training on massive corpora and fine-tuning for a specific downstream purpose, these models can acquire accurate language representations.

Our Inline Citation dataset includes scientific data containing science-specific terminology. Because of that, we decided to encode texts for the classification task using the BERT architecture that has been pre-trained on scientific texts, i.e. the SciBERT model (Beltagy et al., 2019). Just like BERT, SciBERT contains 30K word-piece tokens, but unlike BERT its vocabulary is pertinent to the scientific area. In the scientific domain, SciBERT outperforms BERT in a variety of tasks (Beltagy et al., 2019) and achieves SOTA performance in multi-class text classification on the SciCite dataset (Cohan et al., 2019).

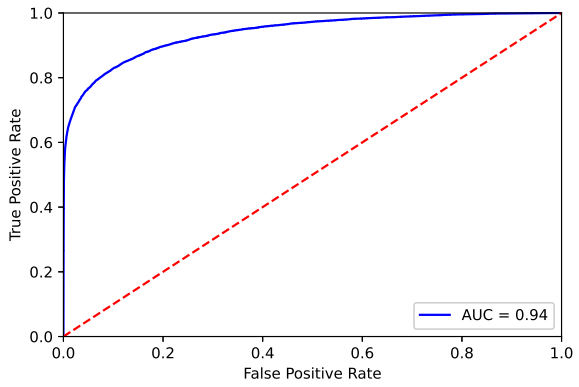

Figure 1: ROC curve testing subset on imbalanced dataset

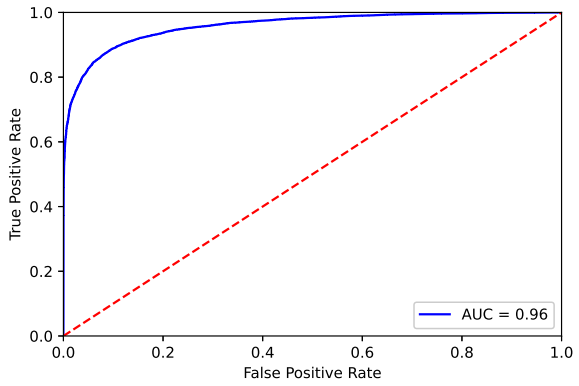

Figure 2: ROC curve testing subset on balanced dataset

It has been demonstrated that fine-tuned uncased SciBERT with SciVocab followed by a linear layer produces the best results for scientific data (Beltagy et al., 2019) or for citation context classification (Maheshwari et al., 2021). Therefore we use this model in each experiment.

4 Fine-tuning SciBERT

In this study, we evaluate both a balanced and an imbalanced dataset. We divided both the balanced and the imbalanced dataset into a training subset (60%), a validation subset (20%) and a test subset (20%), resulting in a "60:20:20" split.

Dataset type	Class	Training subset	Validation subset	Testing subset
Balanced	Contains citation	32831	10957	10947
	No citation	36085	12015	12025
Imbalanced	Contains citation	32739	11049	10947
	No citation	214455	71350	71452

Table 2: Dataset split

The split was then modified so that the proportion of positive (sentences containing a citation) to negative (sentences not containing a citation) texts in each subset would not be altered following the split (see Table 2). Next, we fine-tuned all SciBERT parameters end-to-end utilizing the training and validation subsets. For fine-tuning, we adhered primarily to the similar design and optimization decisions utilized in articles (Beltagy et al., 2019; Devlin et al., 2018). We used the ReLu activation function (Agarap, 2018) in linear one-layer feed-forward classifier which inputs the last hidden state of the [CLS] token. In other words, this last hidden state of the [CLS] token is utilized as the sequence's features to feed the classifier.

We experimented with numerous hyperparameters for fine-tuning with both datasets. We fine-tuned for 2 to 5 epochs using batch size of 16, 32 or 50 and learning rate of 5e-5, 5e-6, 1e-5 or 2e-5, with a dropout of 0.1 or without dropout. We optimized cross-entropy loss with the assistance of the AdamW optimizer (Kingma and Ba, 2014). The best results were obtained when the models were fine-tuned for 2 epochs with a batch size of 50 samples and a learning rate of 5e-5 without dropout, followed by a linear warmup and linear decay (Devlin et al., 2018); this was the case for both the balanced and the imbalanced dataset. We used softmax to determine probabilities for predictions, with a threshold of 0.7 proving optimal, meaning that sentences with a calculated probability greater than 0.7 are predicted positive, i.e. are predicted to contain an inline citation.

5 Discussion

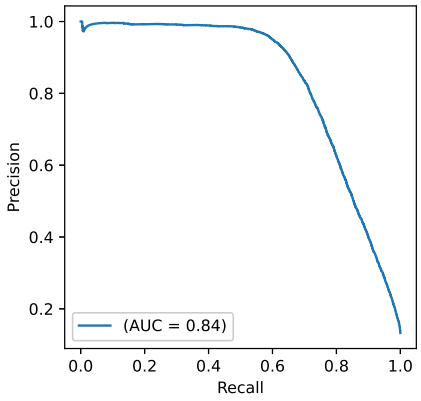

Figure 3: PR curve testing subset on imbalanced dataset

Citation prediction				
Approach	**Precision**	**Recall**	**F1 score**	**Accuracy**
Balanced SciBERT validation	0.93	0.84	0.89	0.90
Balanced SciBERT testing	0.93	0.84	0.88	0.89
Imbalanced SciBERT validation	0.92	0.63	0.75	0.94
Imbalanced SciBERT testing	0.92	0.64	0.75	0.94

Table 3: Prediction results

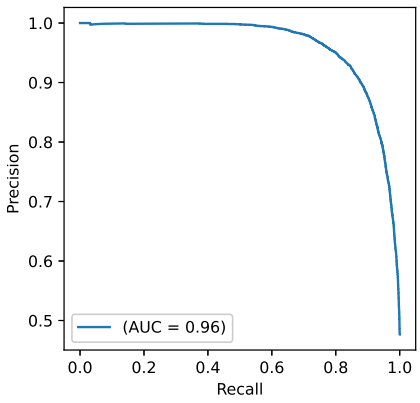

Figure 4: PR curve testing subset on balanced dataset

We report the results of our two experiments in Table 3. Performing classification tasks using imbalanced datasets poses multiple challenges, the most prominent being the bias towards the most represented class (He and Garcia, 2009), i.e. the classification algorithm has a tendency to fail to correctly identify the least represented class. This is a phenomenon that we noticed in our study as well. Balancing the dataset by undersampling helped to significantly reduce this bias, increasing the recall of the least represented class (=sentences containing an inline citation) from 0.63 to 0.84.

Since we used both balanced and imbalanced datasets, useful performance indicators include the Area under the Curve AUC for the precision-recall curve PR or the Receiver Operating Characteristic curve ROC (Bradley, 1997; Hanley and McNeil, 1982). Figure 1 and figure 2 reveal that the ROC curves are nearly comparable in both datasets, with the imbalanced dataset having a slightly lower AUC value of 0.94 against that of 0.96 for the balanced dataset. For imbalanced data, however, a PR plot is advised (Sun et al., 2009; Gu et al., 2009); our PR plots are depicted in figure 3 and 4. The imbalanced dataset's PR curve follows a different path than the balanced dataset's PR curve, which is also reflected in its considerably lower AUC value (=0.84) compared to that of the balanced dataset (=0.96).

6 Conclusion

The goal of this paper was to determine how effective binary classification models can be at predicting whether or not sentences in academic articles should contain an inline citation. To that end, we used regular expressions to identify inline citations in published research papers, and then created a dataset composed of 411k sentences, where approximately 54k contained inline citations. We then ran a fine-tuned SciBERT classifier on both a balanced and imbalanced dataset, achieving an overall accuracy of over 0.92. This result shows that language patterns alone could be used to predict the presence of inline citations in academic text with a reasonable degree of accuracy. We presented the problem as a binary classification task on a sentence level, i.e. we only considered the target sentence and did not consider the context in which the given sentence appeared, for example by also looking at the sentence appearing before and the sentence appearing after the target sentence. The sentences contained in the Inline Citation Dataset however are all sequential: they come in the same sequence as they were found in the original paper. This means that information on the context in which a given target sentence appears is already available in our dataset. This paves the path for further experiments that take contextual sentential information into account, such as training a transformer that tries to predict in which position inline citations should appear, or even open the doors towards experimenting with techniques that could be used to identify the scope of a reference and use the full scope for the classification part as well.

References

Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Andrew P Bradley. 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159.

Joan G. Carson, Nancy D. Chase, Sandra U. Gibson, and Marian F. Hargrove. 1992. Literacy demands of the undergraduate curriculum. *Literacy Research and Instruction*, 31(4):25–50.

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

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. *arXiv preprint arXiv:1804.05685*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Qiong Gu, Li Zhu, and Zhihua Cai. 2009. Evaluation measures of the classification performance of imbalanced data sets. In *Computational Intelligence and Intelligent Systems: 4th International Symposium, ISICA 2009, Huangshi, China, October 23-25, 2009. Proceedings 4*, pages 461–471. Springer.

James A Hanley and Barbara J McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

Haibo He and Edwardo A. Garcia. 2009. https://doi.org/10.1109/TKDE.2008.239 Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Eleonora Kohl Kerstin. 2012. https://doi.org/10.3402/rlt.v19i3.7611 Fostering academic competence or putting students under general suspicion? voluntary plagiarism check of academic papers by means of a web-based plagiarism detection system. *Research in Learning Technology*, 19.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, et al. 2021. Datasets: A community library for natural language processing. *arXiv preprint arXiv:2109.02846*.

Himanshu Maheshwari, Bhavyajeet Singh, and Vasudeva Varma. 2021. Scibert sentence representation for citation context classification. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 130–133.

Alastair Pennycook. 1996. Borrowing others' words: Text, ownership, memory, and plagiarism. *TESOL quarterly*, 30(2):201–230.

Aixin Sun, Ee-Peng Lim, and Ying Liu. 2009. On strategies for imbalanced text classification using svm: A comparative study. *Decision Support Systems*, 48(1):191–201.

Iris Vardi. 2012. Developing students' referencing skills: a matter of plagiarism, punishment and morality or of learning to write critically? *Higher Education Research Development*, 31(6):921–930.

Vijay Viswanathan, Graham Neubig, and Pengfei Liu. 2021. Citationie: Leveraging the citation graph for scientific information extraction.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander Fabbri, Irene Li, Dan Friedman, and Dragomir Radev. 2019. ScisummNet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of AAAI 2019*.