EPiC Series in Computing
Volume 85, 2022, Pages 129–145
Proceedings of 43rd Conference of the South African Institute of Computer Scientists and Information Technologists

Sentiment and intent classification of in-text citations using BERT

Ruan Visser[1,2] and Marcel Dunaiski[1,2]

[1] Department of Computer Science
Stellenbosch University, Stellenbosch, South Africa
[2] School for Data Science and Computational Thinking
Stellenbosch University, Stellenbosch, South Africa
marceldunaiski@sun.ac.za, ruanvisser101@gmail.com

Abstract

Methods such as the *h*-index and the journal impact factor are commonly used by the scientific community to quantify the quality or impact of research output. These methods rely primarily on citation frequency without taking the context of citations into consideration. Furthermore, these methods weigh each citation equally ignoring valuable citation characteristics, such as citation intent and sentiment. The correct classification of citation intents and sentiments can therefore be used to further improve scientometric impact metrics.

In this paper we evaluate BERT for intent and sentiment classification of in-text citations of articles contained in the database of the Association for Computing Machinery (ACM) library. We analyse various BERT models which are fine-tuned with appropriately labelled datasets for citation sentiment classification and citation intent classification.

Our results show that BERT can be used effectively to classify in-text citations. We also find that shorter citation context ranges can significantly improve their classification. Lastly, we also evaluate these models with a manually annotated test dataset for sentiment classification and find that BERT-cased and SciBERT-cased perform the best.

1 Introduction

Scientometrics plays an important role in academia. It helps various research communities evaluate scientific work and allocate resources effectively [15]. In scientometrics, citations form the basis for most prevalent impact indicators used to evaluate academic entities [15], such as the Impact Factor for journals or the *h*-index for authors [16, 19]. Similar to these indicators, most other common metrics treat all citations equally and ignore valuable citation characteristics, such as the sentiment, importance or intent of each citation, even though such characteristics could be leveraged to gain further insight into the scientific consensus on various topics and entities. Metrics that rely primarily on citation counts have frequently been criticised. Moravcsik and Murugesan [31] for instance find that one third of references cited were redundant. Additionally, Simkin and Roychowdhury [36] estimate that researchers only read one fifth of the work they cite, with the majority of citations being copied from other papers' reference lists.

Abu-Jbara et al. [1] state that the number of citations an academic paper receives is not a sufficient evaluation of its quality but rather measures its popularity and researchers' interest in it. Disputed papers, or papers with fabricated experiments, have received many citations [1]. For example, Hwang Woosuk's fraudulent papers [22, 23] on stem cell cloning received almost 200 citations after it was found that his research was dishonest, with the vast majority of these citations being negative [20, 1]. Weighting references by their sentiment can lead to more refined and fairer citation metrics as well as help identify and distinguish between good, bad, or even fraudulent papers.

Similar to citation metrics, in-text citation analysis can also augment other scientometric applications that are based on citation analysis. For example, the ability to distinguish criticism from acclamation and important citations from peripheral mentions can improve applications such as mapping the landscapes of scholarly disciplines, measuring knowledge transfer across domains [14], and improve read recommendations.

Previous attempts to classify in-text citations had significant drawbacks such as the need to add manually annotated features, only using small datasets, or were limited by the available computing resources at the time [40, 1, 18]. However, in recent years natural language processing (NLP) methods have advanced, with deep learning models such as BERT, GPT and ELMO [12, 34, 33] outperforming feature based models in most NLP tasks [39]. Similarly, citation classification has rapidly improved where models such as BERT and ELMO [6, 9] now outperform techniques such as Support Vector Machines (SVM) and Random Forests. Most recent studies on citation classification have focused on classifying citation intent, however, these methods can also be applied to classifying in-text citations according to sentiment or importance.

In this paper we use various BERT models to classify citations with regard to both intent and sentiment. We use the SciCite dataset [9] to train our intent classification models and the Citation Sentiment Corpus [5] to train our sentiment classification models. When classifying citation sentiment with BERT models, the major difficulty is that these models are unable to focus on the sentiment conveyed towards a specific citation within the given text. For this reason we explore another approach to sentiment analysis, namely aspect-based sentiment analysis (ABSA). ABSA aims to evaluate sentiment towards a specific entity or topic within a given text rather than the text itself. Lastly, we evaluate the BERT models which are fine-tuned for sentiment classification on a manually annotated testset of 97 in-text citations from the ACM database.

We give background information on the models used in this paper in Section 2. Section 3 and Section 4 describe the used datasets and methodology, respectively. Lastly, we present and discuss the results in Section 5.

2 Background

2.1 Citation Sentiment

Sentiment analysis is the process of computationally detecting and classifying views conveyed within a text [38]. Extensive research has been conducted in the field of sentiment analysis over the years. Most of this research has been done in general domains such as newspaper sections, product reviews, or social media posts. Implementations using sentiment analysis in these domains have produced good results with many recent models reporting F1 scores greater that 95% [45, 44, 41].

Sentiment analysis implementations within the field of scientific literature, however, have received comparatively little focus and have not been as successful [5]. Athar [5] obtained a macro F1 score of 76% on a custom citation sentiment corpus. Jochim and Schultze [25] use a deep learning model, pretrained on general domains including book and DVD reviews, and obtained a macro F1 score of 54%, which resulted in a 3% improvement when not pretraining with general domain data. According to Athar [5], there are a number of factors that complicate sentiment analysis when applied to scientific literature compared to other domains:

- Sentiment in scientific literature is often implicit, hidden or obfuscated, in particular when negative sentiment towards a citation is conveyed [5].

- Citation contexts often contain science-specific nomenclature and technical terms that carry sentiment and rarely occur in other domains (i.e., state-of-the-art or overfit).

- Citation contexts have varying length, ranging from a single sentence to multiple paragraphs.

- Multiple distinct citations can occur within a single context (and even a single sentence). This is of particular concern when using models that cannot focus on a particular aspect within a text, such as BERT.

- The overwhelming majority of citations have a neutral sentiment. Consequently, many classification models perform poorly when tasked with classifying non-neutral citations.

2.2 Citation intent

A citation can fulfil various roles within a paper. Some citations show explicit use of a tool or method while others are simply used to acknowledge earlier work [9]. Most prominent citation classification models and datasets focus on citation intent, rather than citation sentiment or citation importance [6, 9, 26].

Sentiment classification problems usually use three categories (positive, negative and neutral). However, intent classification problems lack a common and consistent classification scheme. The number of categories range from only 3 to 35 [9, 17]. Cohan et al. [9] argue that some intent categories within fine-grain intent schemes only apply to very few citations. Consequently, it is often challenging to gain insight into their impact. Furthermore, as most citation intent datasets contain less than 2 000 citations, most models struggle to accurately predict rare classes.

Citation intent classification models have been implemented with a fair degree of success. For example, Cohan et al. obtained a macro F1 score of 84% and Beltagy et al. [6] obtained a macro F1 score of 85% using 3 classification categories. The accuracy of citation intent classification can suffer due to similar citation characteristics that hinder sentiment analysis. Two of these complications are the dynamic length of a citation context, and the context overlap, where multiple citations occur within a single context. However, different to citation sentiment classification, citation intent classification requires less context [5, 40]. Both of these complications are therefore less problematic when a shorter context is used to classify citation intent.

2.3 Feature Based Methods

Zhu et al. [48] state that the current citation system is not an adequate method to distinguish the importance of literature, claiming that "not all citations are created equal". The authors created a list of "intuitively attractive" features. However, when testing these features they found that only a few can be effectively used to classify citations. They found that the number of times a paper is referenced (within a citing paper), and the similarity between the citation context and the cited paper's abstract were some of the best predictors for their classification.

Jha et al. [24] propose a more NLP focused solution for intent classification and also introduce a sentiment variable which is extended in our work as a property of a citation. Jha et al. tested different Machine Learning methodologies and found that a SVM approach works best for their citation classification, which was subsequently confirmed by Zhu et al. [48]. When reviewing model performance Jha et al. [24] obtained a macro F1 score of 58%, improving on the 42% obtained by Zhu et al. [48], when categorising a citation as either influential or not.

Teufel et al. [40] found that the number of categories can have a significant impact on model performance. They achieved an F1 score performance improvement from 57% to 71% when they reduced the number of categories from 12 to 3.

2.4 Deep Learning Methods

In recent years significant improvements have been achieved in the field of natural language processing (NLP), with Deep Learning Models outperforming more traditional feature-based models in sentence classification tasks [3]. Consequently, there have been improvements in the field of citation classification. For instance, the BiLSTM-Attention ELMO implementation of Cohan et al. [9], tested on the ACL-ARC database [8], outperforms Jurgens et al.'s Random Forest classifier [27] with a 13% increase in F1 macro score. In contrast to the Jurgens et al.'s model, the model proposed by Cohan et al. [9] does not make use of external linguistic resources nor does it require hand-engineered features. Instead, Cohan et al.'s model makes use of a strategy called structural scaffolding, which utilizes sub-tasks to pretrain the models. These sub-tasks enabled Cohen et al. to improve their model's performance from a macro F1 score of 54% to 67% when tested on the ACL-ARC dataset.

Another deep learning model, BERT [12], has been adapted by Beltagy et al. [6] to perform citation intent classification. They pretrained BERT on a large scientific corpus instead of the general corpus on which the original BERT was trained. Consequently, their model outperforms the original BERT when tasked with intent classification of scientific citations.

3 Data sets

Our main objective is to classify citations from the ACM dataset. Since deep learning models require a large training corpus we use two external datasets to train our BERT models. We use the Citation Sentiment Corpus created by Athar [5] for citation sentiment models and the SciCite dataset created by Cohan et al. [9] for citation intent models.

3.1 Citation Sentiment Corpus

The Citation sentiment corpus contains 8 736 in-text citations each manually annotated according to sentiment. This dataset classifies a citation as positive, neutral or negative. Citations are classified as either positive or negative only if there are polar phrases associated with the cited paper, in contrast to other papers such as Teufel et al. [40] which consider a citation as positive according to its intent.

Table 1: Examples of polar phrases found in the Citation Sentiment Corpus.

Positive Phrases	Negative Phrases
appealing	daunting
straightforward	complicated
improve the performance	degrade
overcome	restrict

Polar phrases are, however, rare in scientific papers since authors are hesitant to use such phrases within their papers and, in particular, when used to criticize other authors. Therefore, the majority of citations within this dataset are neutral, 9.5% are positive, and only 3% are negative. Each instance within the Citation Sentiment Corpus is in the following format:

```
C96-1036:::A92-1018:::o:::"... N-gram class models (Brown et al. , 1992)
and Ergodic Hidden Markov Models (Kuhn el, al. , 1994) were proposed and
used in applications such as syntactic class (POS) tagging for English
(Cutting et al. , 1992), ..."
```

In the above example "`C96-1036`" is the source paper identifier, "`A92-1018`" is the cited papers' identifier, "`o`" is the labelled sentiment, and lastly "`... N-gram class models ...`" is the citation context. In this example it is unclear which citation to focus on which renders the use of an aspect-based model impractical. Fortunately, the implementation by Athar [5] is open source and available on Github together with a test dataset where the specific citations are marked. In this dataset "<CIT>" is used to mark the citations in question and "<OTH>" for all other citations. We use this testset to identify specific aspects in citation contexts. Furthermore, we use Part of Speech tagging to improve aspect extraction prior to classification. The details of these methods are further described in Figure 3.

3.2 SciCite

In the SciCite dataset each citation is classified as either (1) background information, (2) method, or (3) result comparison. Other citation intent datasets with fine-grained classes often only contain few elements per class which makes them impractical to use [9]. SciCite contains more the 11 000 citations with most citations classified as background information. Figure 1 shows the class distribution of the annotated citations in the SciCite dataset.

3.3 ACM dataset

The Association for Computing Machinery (ACM) dataset [2] used in this paper contains full-text papers from the ACM digital library published between 1950 and 2015. In addition to the fulltexts this data set includes internal cross-references between papers. We annotated 97 in-text citations according to both the citation sentiment and the citation function. Each annotation was coded by three annotators, with the agreement between the annotators determined by using the kappa (κ) coefficient [10]. For citation sentiment the κ score is 0.69 while for intent classification the agreement is 0.43. κ scores between 0.61 and 0.8 convey a substantial level of

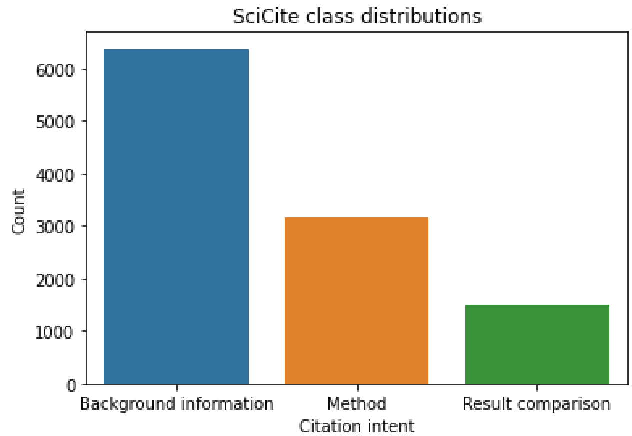


Figure 1: Class distribution of annotated citations in the SciCite dataset used for intent classification.

agreement [29]. Due to the low agreement of 0.43 between the annotators when classifying citation intent, the models predicting intent were not evaluated using the ACM testset. Therefore, we only use the 97 labelled citations as a testset to determine how well each sentiment model generalizes when given in-text citations outside of the training corpus.

4 Methodology

4.1 Models

We use various BERT models to determine both intent and sentiment of in-text citations contexts.

4.1.1 BERT

BERT is based on a multi-layer bidirectional transformer as opposed to the conventional unidirectional language modeling applied by models such as ELMO (which uses two unidirectional Transformers). BERT is trained by predicting both randomly masked tokens and whether two sentences follow one another. For tokenization BERT uses WordPiece [43], which constructs BERT's vocabulary to include the most commonly used words and word-pieces. We use BERT-Base which has 12 layers and 768 hidden dimensions [12]. BERT-Base can be either case sensitive (BERT-Base-cased) or not (BERT-Base-uncased). We evaluate both cased and uncased BERT-Base models.

4.1.2 RoBERTa

The RoBERTa model was pretrained on a significantly larger corpus than BERT. Additionally, it also features some architectural changes such as dynamic masking, full sentence training, and training in large mini-batches. Similar to GPT-2, RoBERTa uses Byte-Pair Encoding (BPE) [35] for tokenization. BPE makes use of subword units, which are derived from statistical analysis of the pretraining corpus [30]. RoBERTa has been found to outperform BERT in many NLP tasks including text classification [30]. We evaluate whether these classification improvements extend to in-text citation classification.

4.1.3 SciBERT

The SciBERT model uses the original BERT code and the same configurations and size as BERT-Base [6]. However, different to BERT, SciBERT is pretrained on the Semantic Scholar Open Research corpus [4] and uses its own vocabulary SciVocab. SciVocab is a WordPiece vocabulary created from a scientific corpus. Due to the differences between scientific text and general domain text, Beltagy et al. [6] found a large disparity (42% difference) between BERT's vocabulary and SciVocab. Since all our datasets contain scientific texts we evaluate SciBERT's performance against the other BERT variants in terms of sentiment identification.

4.1.4 XLNet

XLNet has a similar architecture to BERT, however, it takes an alternative approach to pretraining. Instead of the auto-encoder strategy used by BERT and most popular transformer models, XLNet uses an auto-regressive pretraining approach. In contrast to BERT which does not take the masked positions into account, XLNet accounts for token positions. This enables it to learn bidirectional contexts while maximizing the expected likelihood across all permutations of the factorization order of a given text [46]. XLNet uses SentencePiece [28] for language independent tokenization.

4.1.5 ABSA-BERT

The specific ABSA model used in this paper is LCF-BERT created by Zeng et al. [47]. This model uses a Local Context Focus (LCF) technique for aspect-based sentiment analysis which utilized a Context features Dynamic Mask (CDM) and a Context features Dynamic Weighted (CDW) layers in order to emphasise the local context.

4.2 Parsing

We use ParsCit [11] to extract in-text citations and references from papers within the ACM dataset. ParsCit is an open source implementation of a reference string parsing package, which uses conditional random fields (CRF) to label reference strings. In addition to CRF, ParsCit also uses a heuristic model which enables it to identify reference strings from plain text and retrieve citation contexts [11]. We selected ParsCit as a citation extraction framework due to its ability to automatically extract citations with context and its support for fulltext papers in text format. Before ParsCit can be used the paper fulltexts must be cleaned and converted to emulate ParsCit's templates[1]. Accordingly, we performed the following steps to extract the citation contexts:

- Add new lines after each 15$^{\text{th}}$ word[2].
- Regularize citation tags and clean fulltexts.
- Extract citations using ParsCit.
- Retrieve citations and contexts from ParsCit's output.

[1] For template examples, see https://github.com/knmnyn/ParsCit/tree/master/test/txt.
[2] This step is needed to emulate ParsCit's templates.

See Figure 2 for the full ACM data processing workflow.

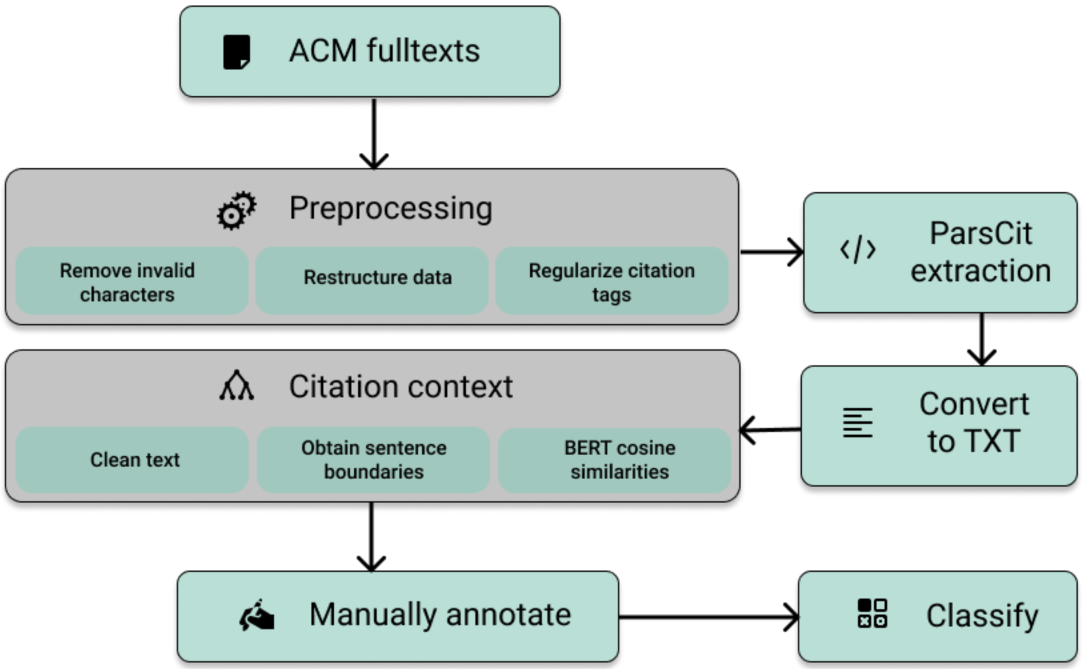


Figure 2: The workflow used in this paper to extract and classify in-text citation from noisy fulltexts within the ACM digital library. Subsequently, the extracted in-text citation contexts were manually annotated according to both sentiment and intent. Lastly, every in-text citation was classified by all models which were fine-tuned on the Citation Sentiment Corpus.

4.3 Citation context

When classifying a citation context no fixed or predefined context scope is available. Some citation contexts are limited to one sentence while others span over several paragraphs. The citation context that ParsCit extracts is static and fixed to 400 characters. Fixed contexts are often difficult to classify since a large amount of the context may be irrelevant or may contain parts of another citation's context. In addition to the previous complications BERT's performance is known to degrade when classifying longer sequences [12, 7]. To mitigate these issues we identify the most relevant sentences within a given context dynamically by performing the following steps:

- Split the sentences using Spacy [21].
- Remove incomplete sentences at the start and end of the context.
- Identify the sentences containing the citation in question.
- Vectorize the sentences with BERT-Base.
- Calculate the cosine similarity between each sentence and the citation sentences.

- Remove sentences according to both cosine similarity and their location in a given context.
- Either remove citation tags or replace them with a generic term.

Sentences were removed according to the following heuristic formula:

$$\frac{consine\_similarity}{index + 1} > 0.075 \tag{1}$$

This formula gives precedence to sentences later in a context to avoid removing sentences with a negative sentiment, which commonly occur after a citation sentence, also known as hedging [13, 48]. We further preprocess contexts by identifying and handling implicit and explicit citation tags (see Figure 3). We define a citation tag as explicit when the citation is acknowledged within a sentence and implicit when the converse is true. We remove implicit tags since they can obstruct the structure of a sentence. Therefore explicit citation tags are replaced by a generic term, i.e. "*this paper*".

When data is prepared for ABSA-BERT the appropriate aspect has to be specified. However, in both the Citation Sentiment Corpus and the ACM dataset the specific cited aspect is unknown. To find the cited aspect we use citation tags in conjunction with Part of Speech tagging to identify the cited aspect. See Figure 3 for aspect identification examples.

4.4 Control Parameters

The SciCite dataset was split into 75% training set, 10% validation set, and 15% testset. To ensure that the models have a sufficient amount of polar examples when training, we increase the training split and reduce the validation split when evaluating the Citation Sentiment Corpus (CSC). Consequently, the CSC was split into 80% training set, 5% validation set, and 15% testset.

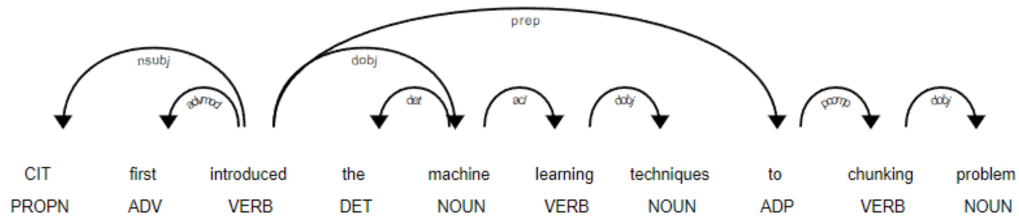
We used 8 epochs and early stopping, with a patience of 3 evaluations during training, and an Adam optimizer to adjust the model weights. We selected a maximum sequence length according to the token length densities shown in Figure 4. As can be seen in the Figure 4, sentences rarely have a sequence length larger that 100. However, as hedging usually occurs later in a context, we selected 128 as the maximum sequence length.

We performed a grid search over the control parameters listed in Table 2 to determine the optimal control parameters for each task-specific model, with each configuration being evaluated with stratified shuffle split 3-fold cross validation. Since smaller learning rates performed better for ABSA-BERT we set the learning rates to $5e^{-6}$, $1e^{-5}$, and $1.5e^{-5}$ when performing grid search for ABSA-BERT. Table 3 lists the optimal control parameters for each model when trained on either CSC or the SciCite dataset.

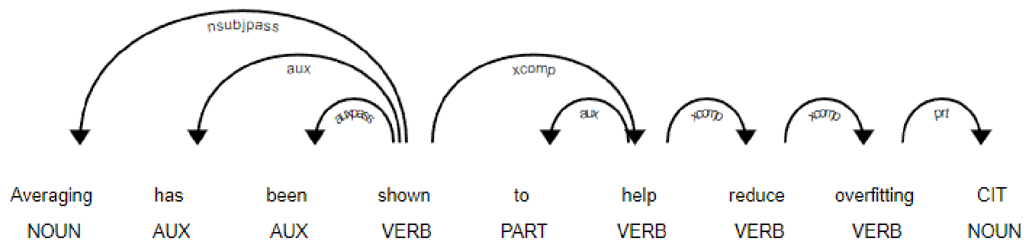
After the optimal control parameters were set, we evaluated each model's best checkpoint on the held-out testsets and the ACM dataset by using F1 macro to evaluate validation performance.

4.5 Implementation

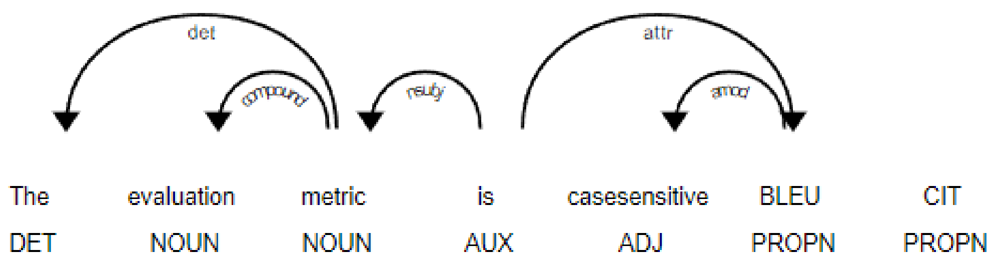
Pytorch [32] and the Transformer library [42] were used to import, train and evaluate BERT, SciBERT and RoBERTa. For aspect-based sentiment analysis we used pyabsa to load, train and evaluate the ABSA-BERT. To ensure consistent results we used a fixed random seed of 42.



(a) In this example the citation is explicitly acknowledged in the context. This can be inferred due to the CIT tag's location at the start of the sentence. Therefore, the given citation tag (CIT) is the aspect.



(b) Since this citation is both at the end of the sentence and the after a verb, the whole sentence is considered to be the aspect.



(c) In this example the word prior to the citation tag is a proper noun. In this case, the citation tag is considered to be the aspect along with all of its children. In this example "casesensitive BLEU" is the cited aspect.

Figure 3: Examples of how cited aspects were identified.

Table 2: Control parameters tested for each model.

Control Parameter	Values
Learning rate	$2e^{-5}$, $3e^{-5}$, $4e^{-5}$
Dropout	0.3, 0.5, 0.7
Batch size	16, 32

4.6 Evaluation Metrics

We use accuracy and F1 macro to evaluate each model. The accuracy metric is the proportion of correct predictions amongst all instances examined [37]. TP is the number of true positives and TN is the number of true negatives predicted by the model while FP is the number of false

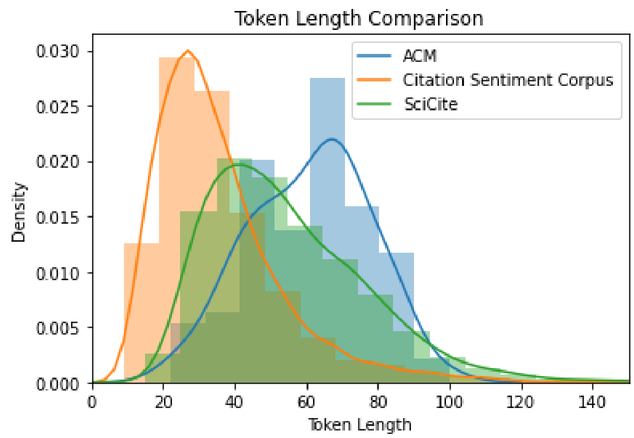


Figure 4: The distribution over citation contexts lengths found in the three datasets used in this paper.

Table 3: The optimal control parameters found for each model on both the Citation Sentiment Corpus (CSC) and SciCite databases.

	CSC			SciCite		
Model	LR	Dropout	Batch size	LR	Dropout	Batch size
BERT-uncased	$3e^{-5}$	0.3	32	$2e^{-5}$	0.3	16
BERT-cased	$2e^{-5}$	0.5	32	$2e^{-5}$	0.3	32
RoBERTa	$2e^{-5}$	0.3	16	$4e^{-5}$	0.3	16
SciBERT-uncased	$3e^{-5}$	0.5	32	$4e^{-5}$	0.3	32
SciBERT-cased	$2e^{-5}$	0.5	32	$2e^{-5}$	0.3	16
XLNet	$2e^{-5}$	0.5	16	$3e^{-5}$	0.7	32
ABSA-BERT	$1e^{-5}$	0.3	16	-	-	-

positives and FN is the number of false negatives:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad (2)$$

In binary classification F1 measures the relationship between the data's positive labels and those given by the classifier [37]. Macro F1 is used in multi-class classification, in which the mean of all classes' F1-scores are calculated. Macro F1 weighs all classes equally regardless of their densities. Therefore, when classifying an unbalanced dataset, macro F1 can be useful to show when a models overfits the majority class. $\overline{Precision}$ and $\overline{Recall}$ are the arithmetic means of the Precision and Recall for each class.

$$F1_{macro} = \frac{2 \times \overline{Precision} \times \overline{Recall}}{\overline{Precision} + \overline{Recall}} \quad (3)$$

Table 4: Sentiment classification result on the Citation Sentiment Corpus (CSC) and the ACM testset, as well as, intent classification result on the SciCite dataset.

	CSC		SciCite		ACM	
Model	F1	Accuracy	F1	Accuracy	F1	Accuracy
BERT-uncased (static)	-	-	-	-	0.56	0.65
BERT-uncased	0.62	0.89	0.83	0.84	0.62	0.69
BERT-cased	0.63	0.88	0.84	0.86	0.76	0.81
RoBERTa	0.61	0.89	0.84	0.85	0.51	0.68
SciBERT-uncased	0.63	0.87	0.85	0.86	0.65	0.72
SciBERT-cased	0.67	0.88	0.84	0.86	0.74	0.78
XLNet	0.60	0.86	0.85	0.87	0.57	0.70
ABSA-BERT	0.69	0.90	-	-	0.72	0.79

5 Results

Table 4 shows the results of the experiments for sentiment analysis using the Citation Sentiment Corpus (CSC) and the ACM testset, and results for experiments for intent classification using the SciCite dataset.

5.1 Citation Sentiment Corpus

The baseline model BERT-uncased achieves an F1 score of 62% on our Citation Sentiment Corpus testset. We find that case sensitivity marginally increases the performance of BERT to 63%. The improvement by including case sensitivity is more pronounced for the SciBERT model. It improved by 4% points to a final score of 67%. We find that XLNet and RoBERTa achieve the worst results with F1 scores of 60% and 61%, respectively. Lastly, we find that ABSA-BERT performs the best with an F1 score of 69%.

The improvement of SciBERT over the baseline BERT model in predicting sentiment within the Citation Sentiment Corpus is expected since SciBERT is specifically pretrained on a scientific document corpus. However, we find that both the XLNet and RoBERTa models, which aim to improve on BERT, perform worse than the baseline BERT-uncased. These results are in contradiction to the results obtained on most classical benchmarks for which RoBERTa and XLNet commonly outperform the original BERT models [30][46].

Since BERT, RoBERTa, SciBERT and XLNet classify the sentiment of the text as a whole and not the sentiment conveyed towards a specific citation, these models seem to be ineffective when classifying a text which contains multiple citations. Consequently, we expect a model such as ABSA-BERT to perform better when the sentiment of a text as a whole differs from the sentiment directed towards a specific citation. In Table 5 we present a few examples in which ABSA-BERT identified an aspect's sentiment where the other models did not.

5.2 SciCite

Confirming the results found by Beltagy et al. [6], we find that SciBERT-uncased performs better than BERT-uncased and SciBERT-cased when classifying citation intent in SciCite. However, the difference in performance between all models is marginal, with the worst performing model obtaining a macro F1 score of 83%, only 2% points worse than the best model. In

Table 5: Selected examples from the Citation Sentiment Corpus. The first column lists the in-text citations as in the Citation Sentiment Corpus. The second column shows the truth sentiments. The predicted sentiments according to BERT-uncased and ABSA-BERT are shown in columns 3 and 4 respectively. Lastly, the citation aspects used by the ABSA-BERT model is given in column 5.

Text	Truth	BERT-uncased	ABSA	Aspect
Our system improves over the latent named entity tagging in <CIT>, from 61 to 87	Negative	Positive	Negative	*explicit*
An alternative method <CIT> makes decisions at the end but has a high computational requirement	Negative	Neutral	Negative	An alternative method
However as discussed in prior arts <CIT> and this paper linguistically informed SCFG is an inadequate model for parallel corpora due to its nature that only allowing child node reorderings	Neutral	Negative	Neutral	prior arts
The time complexity of the CKY-based binarization algorithm is n3 which is higher than that of the linear binarization such as the synchronous binarization <CIT>	Neutral	Negative	Negative	the synchronous binarization
For a full derivation of the modified updates and for quite technical convergence proofs see <CIT>	Positive	Neutral	Neutral	*explicit*
Our experiments on the Canadian Hansards show that our unsupervised technique is significantly more effective than picking seeds by hand <CIT> which in turn is known to rival supervised methods	Negative	Positive	Negative	hand
So unlike some other studies <CIT> we used manually annotated alignments instead of automatically generated ones	Neutral	Negative	Neutral	some other studies

contrast to the results for sentiment classification, we find that XLNet performed best for intent classification with a F1 score of 85% and an accuracy of 87%. ABSA-BERT was not tested on this dataset as there are no citation tags found within the SciCite dataset.

5.3 ACM Corpus

When evaluating the sentiment classification results on the ACM corpus we find that a shorter dynamic context improves the results for BERT-uncased, with an improvement of 6% points in F1 results. Similar to the result obtained in the Citation Sentiment Corpus, RoBERTa and XLNet both perform the worst out of all models evaluated. The case sensitive SciBERT and BERT perform better than their uncased counterparts.

Although ABSA-BERT performs best in the Citation Sentiment Corpus it did not generalize well when evaluated on the ACM dataset when compared to SciBERT-cased and BERT-cased. This could be attributed to the longer sequence lengths found in this corpus compared to the Citation Sentiment Corpus for which the ABSA-BERT model is expected to perform worse.

6 Conclusion and Future Work

We showed that BERT can be used effectively to classify both citation intent and sentiment of in-text citations. We illustrated how dynamically determining citation context sizes can significantly improve citation classification performance. Furthermore, we found that case sensitive models perform better than their uncased counterparts when classifying the sentiment of citation contexts. Lastly, we found that BERT-cased and SciBERT-cased generalize best on our manually annotated ACM testset for sentiment classification.

In this paper we only evaluated a single ABSA model. Future research is required to compare the performance of different ABSA model variations, as well as more extensive parameter fine-tuning. We suggest this is done with both additional in-text citations and more general domain aspect-based datasets. Furthermore, we suggest that further in-depth error analysis is performed to better understand the various models' behaviours and the discrepancies between the obtained results.

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. Association for Computational Linguistics, 2013.

[2] Inc. ACM. Acm digital library, 2016.

[3] Basemah Alshemali and Jugal Kalita. Improving the reliability of deep neural networks in nlp: A review. *Knowledge-Based Systems*, 191:105210, 2020.

[4] Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, ..., and Oren Etzioni. Construction of the literature graph in semantic scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 84–91. Association for Computational Linguistics, 2018.

[5] Awais Athar. Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 Student Session*, HLT-SS '11, page 81–87, USA, 2011. Association for Computational Linguistics.

[6] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics.

[7] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *ArXiv*, abs/2004.05150, 2020.

[8] Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, ..., Min-Yen Kan, and Yee Fan Tan. The ACL Anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 2008. European Language Resources Association (ELRA).

[9] Arman Cohan, Waleed Ammar, Madeleine Zuylen, and Field Cady. 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, 2019.

[10] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[11] Isaac G. Councill, Clyde L. Giles, and Min-Yen Kan. Parscit: An open-source crf reference string parsing package. In *Proceedings of the Language Resources and Evaluation Conference (LREC 08)*, 2008.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019.

[13] Chrysanne Di Marco and Robert Mercer. Hedging in scientific articles as a means of classifying citations. *AAAI Spring Symposium - Technical Report*, 01 2004.

[14] Ying Ding, Guo Freeman, Tamy Chambers, Min Song, Xiaolong Wang, and Chengxiang Zhai. Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology*, 65:1820–1833, 2014.

[15] Marcel Dunaiski, Willem Visser, and Jaco Geldenhuys. Evaluating paper and author ranking algorithms using impact and contribution awards. *Journal of Informetrics*, 10(2):392–407, 2016.

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

[17] Mark A. Garzone. *Automated Classification of Citations Using Linguistic Semantic Grammars*. Canadian theses on microfiche. University of Western Ontario, 1997.

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

[19] Jorge E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572, 2005.

[20] Sungook Hong. The hwang scandal that "shook the world of science". *East Asian Science, Technology and Society: An International Journal*, 2(1):1–7, 2008.

[21] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. Accessed: 2020-09-30, 2017.

[22] Woo Suk Hwang, Sung Il Roh, Byeong Chun Lee, Sung Keun Kang, Dae Kee Kwon, Sue Kim, ..., and Gerald Schatten. Patient-specific embryonic stem cells derived from human scnt blastocysts. *Science*, 308(5729):1777–1783, 2005.

[23] Woo Suk Hwang, Young June Ryu, Jong Hyuk Park, Eul Soon Park, Eu Gene Lee, Ja Min Koo, ..., and Shin Yong Moon. Evidence of a pluripotent human embryonic stem cell line derived from a cloned blastocyst. *Science*, 303(5664):1669–1674, 2004.

[24] Rahul Jha, Amjad Abu-Jbara, Vahed Qazvinian, and Dragomir R. Radev. Nlp-driven citation analysis for scientometrics. *Natural Language Engineering*, 23:93 – 130, 2016.

[25] Charles Jochim and Hinrich Schütze. Improving citation polarity classification with product reviews. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 42–48, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[26] 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.

[27] 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.

[28] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, 2018. Association for Computational Linguistics.

[29] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

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

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

[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, ..., and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[33] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *NAACL*, pages 2227–2237. Association for Computational Linguistics, 2018.

[34] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *ArXiv*, 2018.

[35] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, 2016. Association for Computational Linguistics.

[36] Mikhail V. Simkin and Vwani P. Roychowdhury. Read before you cite! *Complex Syst.*, 14:269–274, 2003.

[37] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009.

[38] Angus Stevenson. *Oxford Dictionary of English*. Oxford University Press, 2010.

[39] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. *CoRR*, abs/1906.02243, 2019.

[40] Simone Teufel, Advaith Siddharthan, and Dan Tidhar. Automatic classification of citation function. In *2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pages 103–110. ACL, 2006.

[41] Tan Thongtan and Tanasanee Phienthrakul. Sentiment classification using document embeddings trained with cosine similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 407–414, Florence, Italy, July 2019. Association for Computational Linguistics.

[42] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, ..., and Alexander M. Rush. 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, October 2020. Association for Computational Linguistics.

[43] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc Le, Mohammad Norouzi, Wolfgang Macherey, ..., and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human

and machine translation. *ArXiv*, 2016.

[44] Qizhe Xie, Zihang Dai, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. *arXiv: Learning*, 2020.

[45] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[46] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237, 2019.

[47] Biqing Zeng, Heng Yang, Ruyang Xu, Wu Zhou, and Xuli Han. Lcf: A local context focus mechanism for aspect-based sentiment classification. *Applied Sciences*, 9:3389, 2019.

[48] Xiao-Dan Zhu, Peter D. Turney, Daniel Lemire, and André Vellino. Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*, 66, 2015.