Citation Context Classification: Critical vs Non-critical

Sonita Te[1], Amira Barhoumi[1], Martin Lentschat[1], Frédérique Bordignon[2,3], Cyril Labbé[1], and François Portet[1]

[1]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
sonita.te@cadt.edu.kh , {firstName.lastName}@univ-grenoble-alpes.fr
[2]Ecole des Ponts, Marne-la-Vallée, France
[3]LISIS, CNRS, INRAE, Univ Gustave Eiffel, Marne-la-Vallée, France
frederique.bordignon@enpc.fr

Abstract

Recently, there have been numerous research in Natural Language Processing on citation analysis in scientific literature. Studies of citation behavior aim at finding how researchers cited a paper in their work. In this paper, we are interested in identifying cited papers that are criticized. Recent research introduces the concept of *Critical citations* which provides a useful theoretical framework, making criticism an important part of scientific progress. Indeed, identifying critics could be a way to spot errors and thus encourage self-correction of science. In this work, we investigate how to automatically classify the critical citation contexts using Natural Language Processing (NLP). Our classification task consists of predicting critical or non-critical labels for citation contexts. For this, we experiment and compare different methods, including rule-based and machine learning methods, to classify critical vs. non-critical citation contexts. Our experiments show that fine-tuning pretrained transformer model *RoBERTa* achieved the highest performance among all systems.

1 Introduction

In scientific papers, citations acknowledge the sources and help the reader to find more information about the citation context. Citations are also an important indicator exploited to identify significant publications in a specific scientific field (Aragón, 2013). They are used for different purposes, e.g. referring to state of the art, to a specific method or result, and they reflect how authors frame their work and this diversity impacts future academics' adoption (Jurgens et al., 2018).

According to Bordignon (2022), the study of critical citation appears to give an applicable theoretical framework, making criticism a vital phenomenon for scientific development. We believe that classifying citation contexts into critical/non-critical categories could be essential to downstream process, such as identifying scientific claims or observing controversial papers.

Bordignon (2022) identifies three different functions for *Critical citation context* : "to criticize," "to compare," and "to question" where :

- "to criticize" function refers when the citing paper points out a weakness or a fault in the cited paper. For instance, *"X1 method did not work well, although they reported 80% accuracy in (Y1 and Y2, 2002)."*
- "to compare" function refers to a link made between two studies with the indication that one research is superior to another, without necessarily including one's own work. One must have the criticizing meaning in the citation contexts. For example, *"(Y1 and Y2, 2008) outperformed (Y3 and Y4, 2007)."*.
- "to question" function refers to a citation made by the citing paper to raise concerns, doubts, and uncertainty about the cited paper. For instance, *"Thus, the full model proposed by Y1 (2002) has remained empirically unproven."*

There have been numerous researches on citation analysis in NLP, with for instance determining citation sentiments (Athar, 2011; Liu, 2017). In addition to citation sentiment, there have been research to define citation function which refers to the specific purpose a citation plays with respect to the citing paper (Bakhti et al., 2018; Jurgens et al., 2016; Pride et al., 2019; Yu et al., 2020). These researches have been conducted to find the real reason behind the citation. Nevertheless, how citation might be utilized to point out criticism and encourage correction have not been studied yet.

Given a set of citation contexts, our work aims at determining critical ones using NLP methods.

Proceedings of the Third Workshop on Scholarly Document Processing, pages 49–53
October 12–17, 2022.

First, we present the construction process of the corpus, which contains citation contexts annotated with critical and non-critical labels. Then, we experiment different methods to classify citation contexts into critical/non-critical labels using our constructed corpus. Indeed, we compare and discuss rule-based methods and machine learning ones.

2 Related Works

In this section, we present different existing works for citation analysis. Some of them are rule-based methods, while others are based on machine learning methods.

Since 2000, several researches on automation citation classification have been using rule-based approaches (Garzone and Mercer, 2000; Nanba et al., 2000; Pham and Hoffmann, 2003). The rule creation process is generally composed of 2 steps. In the first step, cue words/phrases are extracted from dataset samples. In the second step, rules are created based on the extracted cue words/phrases. These rules are the bases to classify citation contexts. For instance, in (Avanço, 2020) a rule-based method is used to identify negative or contradictory citation contexts. The authors built *CitaNeg* corpus (Table 1) and created functions (linguistic patterns) grouped by category: 13 functions for weakness category (WF), 5 functions for compare category (CF), 4 functions for background category (BF), 6 functions for hedges category (HF) and 14 for additional category (GF). However, only WF and CF categories were used for evaluation giving a precision of 0.72 and a recall of 0.69.

More recently, several approaches relying on machine learning have been proposed. For example, Teufel et al. (2006) used IBk, a form of K-Nearest Neighbor (kNN), to classify citation contexts into 4 polarities (Weakness, Positive, Contrast and Neutral) and obtained an f1-score of 0.61 using *Athar* corpus (Table 1). Jurgens et al., 2016, 2018 introduced a representative corpus containing nearly 2 000 citations annotated with 6 labels (background, motivation, extension, use, contrast or future) and reached an f1-score of 0.53 with a Random Forest classifier on their data and a portion of CFC (cf. Table 1). Raza et al. (2019) conducted citation sentimental analysis and citation function analysis by experimenting six machine learning models (Naïve-Bayes, Support Vector Machine, Logistic Regression, Decision Tree, K-Nearest Neighbors and Random Forest). Using *CFC* corpus, the SVM model gave the best performance with an f1-score of 0.88. Using deep learning techniques, Nicholson et al. (2021) developed a *smart citation index* called *scite*, which classifies citations based on their contexts. It indicates whether the context mentions, supports or contrasts the citation. *Scite* is trained on more than 880 million labeled citation contexts, but this data is proprietary and not publicly available. Recently, Karim et al. (2022) evaluated convolutional neural network (CNN) for citation sentiment analysis using different pre-trained word embeddings such as fastText and GloVe. With GloVe embeddings, their CNN model obtained a precision of 0.94 on the *Athar* corpus. Finally, Visser and Dunaiski (2022) used the pre-trained transformer model RoBERTa for citation sentimental analysis and obtained an accuracy of 0.89 on *Athar*.

Table 1 regroups different existing citation context corpora available to the community.

Type	**Name**	**Size**
Citation sentiment	Athar (Athar, 2011)	8 736
	Liu (Liu, 2017)	3 581
	CitaNeg (Avanço, 2020)	19 309
	Critical corpus [1]	1 690
Citation function	CFC (Teufel et al., 2006)	2 829
	Concit (Hernández and Gómez, 2015)	2 195
	IMS (Jochim and Schütze, 2012)	2 008
	DFKI (Dong and Schäfer, 2011)	1 768

Table 1: Available citation context corpora (the Size column contains the number of citation contexts).

3 Experimental setup

We present our methods used for critical/non critical classification in section 3.1. Then, we describe our corpus in section 3.2.

3.1 Methods

We experimented different classification methods to predict critical/non-critical classes. Two rule-based methods (RB and RB+) and 3 machine learning ones (LR, CNN and F-Roberta) were tested.

[1] Critical corpus is provided by LISIS and LIGM and will be published soon

- **RB** represents the rule-based method proposed by (Avanço, 2020). It is considered as a baseline in this work.
- **RB+** represents the improved version of RB method after analyzing and selecting only the rule functions corresponding to the definition of critical citation in Bordignon (2022).
- **LR** refers to Logistic Regression using Tf-Idf for n-grams in range of 1 and 3 grams
- **CNN** represents an inspiration of Karim et al. (2022) using CNN with Glove embeddings.
- **F-RoBERTa** represents a Fine-tuning RoBERTa (Visser and Dunaiski, 2022).In this model, we assigned the class weights of the training set to the model during training in order to deal with imbalance dataset[2].

3.2 Corpus

In order to build a corpus containing critical and non-critical citation contexts, we used available existing annotated datasets presented in Table 1. For the critical class, we used *Critical* corpus which contains 1 690 critical citation contexts. The non-critical citation contexts have been selected from *CitaNeg* dataset based on the definitions of citation functions. In fact, we kept only citation functions that don't contain critical meaning in their definitions. Our final corpus contains 2 413 citation contexts: 1 464 critical citation contexts and 949 non-critical citation contexts. The dataset was randomly split into training and test sets of 75% and 25%, respectively. Table 4 shows the number of citation contexts in the training and test sets.

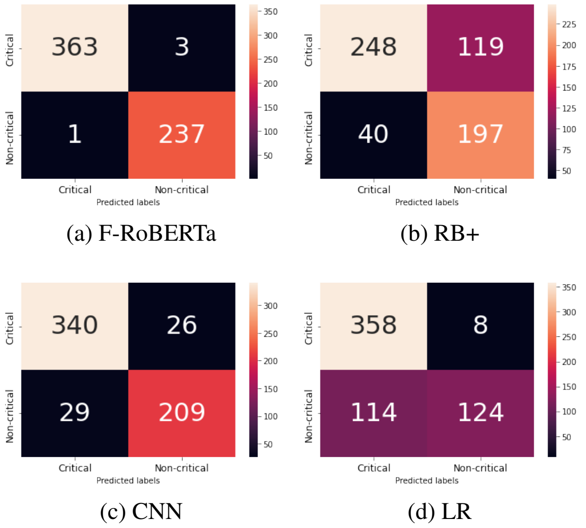
	Train	**Test**	**Overall**
Critical	1098	366	2643
Non-critical	711	238	

Table 2: Train and test sets with numbers of citation contexts

4 Results and Discussion

Table 3 exhibits the performances of the models on the test set. It can be seen that *F-RoBERTa* outperformed all other models. Foremost, we observe that machine learning based approaches systematically outperform rule-based ones.

The confusion matrix in Figure 1 shows that the rule-based system *RB+* has some difficulties in the prediction of critical class (119 are misclassified among 366 critical citation contexts). To improve the quality of the *RB+* system, we need to add more rules to identify critical citation contexts. For instance, we could analyze in depth the grammar or cue words/phrases to define more patterns for critical citation contexts. We could also analyze the concept "to question" of critical citation context that has not been taken into account by our rule-based system *RB+* yet.



(a) F-RoBERTa

(b) RB+

(c) CNN

(d) LR

Figure 1: Confusion matrix of the methods

If we take a look at confusion matrix of *LR* and *CNN* systems in Figure 1d and Figure 1c respectively, the number of misclassified non-critical examples is greater than the number of misclassified critical ones. It could be explained by imbalanced training set. Indeed, critical class represents around 60% of the training set. Being aware of imbalanced training set, we might enhance *CNN* and *LR* performances by assigning class weights while training. However, the *CNN* model does not exhibits a strong bias towards a particular class, so it is likely that a class weighting strategy would have a marginal impact on the performance.

F-RoBERTa (Figure 1a) predicts well non-critical examples, only 1 non-critical and 3 critical examples are misclassified. This could be explained by the class weights' assignment while training *F-RoBERTa* in order to deal with class imbalance. To go further, we analysed these 4 misclassified examples. One of them is *"It is based on modal logic and owes much to the work of Blackburn 1994."* has been classified critical while it should not. If we check out the linguistic aspect of the citation context above, the use of the

[2] We tested RoBERTa model without/with class weights. We reported in this paper the best results obtained with RobERTa with class weights assignment (F-RoBERTa)

Approach	**Method**	**Accuracy** (%)	**Precision** (%)	**Recall** (%)	**F1-score** (%)
Rule-based	RB	73.50	86.31	67.02	75.46
	RB+	74.33	86.30	68.66	76.47
Machine Learning	LR	80.00	75.84	97.81	85.44
	CNN	91.00	88.93	87.81	88.37
	F-RoBERTa	**98.84**	**98.73**	**98.31**	**98.52**

Table 3: Evaluation results of experimented methods

word *"owes"* may reflect critical citation aspect. But, we still can argue if *"owes"* here did not be used to criticize the cited paper, it seems like an incomplete context. In this case, we might need more investigation of corpus. The misclassified critical citation contexts are reported in Table 4. Such miss-classification by *F-RoBERTa* could be explained by the existence of positive and negative words in the same citation context. For example in *Doc_2*, *"perform very well"* is positive and *"dramatically fails"* is negative. To go further, we will use attention mechanism to determine relevant words participating in the prediction.

Critical citation contexts
Doc_1: The morphological processing in PairClass (Minnen et al., 2001) is more sophisticated than in Turney (2006).
Doc_2: In particular, we showed that using a general purpose machine translation (MT) system such as SYSTRAN, or a general purpose parallel corpus - both of which perform very well for news stories (Peters, 2003) - dramatically fails in the medical domain.
Doc_3: In particular, these problems affect the processing of predicate argument structures annotated in PropBank (Kingsbury and Palmer, 2002) or FrameNet (Fillmore, 1982).

Table 4: Misclassified critical examples by *F-RoBERTa*

5 Conclusion and Future work

In this paper, we were interested in identifying critical citation contexts in scientific papers. We proposed and tested five methods for citation context classification into critical/non-critical labels. The methods *RB* and *RB+* were rule-based. The three others, *LR*, *CNN* and *F-RoBERTa*, were machine learning based. We also built a corpus to evaluate and compare these methods. Our task-specific corpus was composed of 2643 citation contexts labeled as being critical or non-critical. Machine learning based systems outperformed rule-based ones. The best system *F-RoBERTa* gave 98.84% of accuracy and 98.52% of F1-score. The performances could be explained by the use of transfer learning in *F-RoBERTa*. Class weight assignment while training might also explain the good accuracy of *F-RoBERTa*'s performance compared to other systems, since our training set was imbalanced.

Some improvements can be made to the proposed systems. In particular, we will assign class weights while model training to solve imbalanced datasets. Moreover, we could operate the data itself (and not the model) to balance the corpus by applying sampling methods either oversampling or undersampling. Dealing with scientific documents, It could be crucial to train our best system *F-RoBERTa*, initially trained on standard corpora, on scientific texts by using for example SciBERT embeddings (Beltagy et al., 2019). Another perspective consists on expanding corpus. To go further, we would extend this work of identifying critical citation contexts in NLP field and study field portability. Indeed, we would identify critical citations in other fields, such as biology or medicine.

Acknowledgments

This work has been undertaken within the *NanoBubbles* project which has received Synergy grant funding from the European Research Council (ERC), within the European Union's Horizon 2020 program, grant agreement no. 951393. It has also been partially supported by MIAI@Grenoble Alpes (ANR-19-P3IA-0003).

References

Alejandro M Aragón. 2013. A measure for the impact of research. *Scientific reports*, 3(1):1–5.

Awais Athar. 2011. Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 Student Session*, pages 81–87, Portland, OR, USA. Association for Computational Linguistics.

Karla Fernanda F. C Avanço. 2020. *Typologie des citations négatives dans les publications scientifiques*.

Khadidja Bakhti, Zhendong Niu, Abdallah Yousif, and Ally Nyamawe. 2018. *Citation Function Classification Based on Ontologies and Convolutional Neural Networks*, pages 105–115.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In *EMNLP*.

Frederique Bordignon. 2022. Critical citations in knowledge construction and citation analysis: from paradox to definition. *Scientometrics*, 127(2):959–972.

Cailing Dong and Ulrich Schäfer. 2011. Ensemble-style self-training on citation classification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 623–631, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Mark Garzone and Robert E. Mercer. 2000. Towards an automated citation classifier. In *Advances in Artificial Intelligence*, pages 337–346, Berlin, Heidelberg. Springer Berlin Heidelberg.

Myriam Hernández and José M. Gómez. 2015. Concitcorpus: Context citation analysis to learn function, polarity and influence.

Charles Jochim and Hinrich Schütze. 2012. Towards a generic and flexible citation classifier based on a faceted classification scheme. In *Proceedings of COLING 2012*, pages 1343–1358, Mumbai, India. The COLING 2012 Organizing Committee.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the Evolution of a Scientific Field through Citation Frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

David Jurgens, Srijan Kumar, Raine Hoover, Daniel A. McFarland, and Dan Jurafsky. 2016. Citation classification for behavioral analysis of a scientific field. *CoRR*, abs/1609.00435.

Musarat Karim, Malik Muhammad Saad Missen, Muhammad Umer, Saima Sadiq, Abdullah Mohamed, and Imran Ashraf. 2022. Citation context analysis using combined feature embedding and deep convolutional neural network model. *Applied Sciences*, 12(6).

Haixia Liu. 2017. Sentiment analysis of citations using word2vec. *CoRR*, abs/1704.00177.

Hidetsugu Nanba, Noriko Kando, and Manabu Okumura. 2000. Classification of research papers using citation links and citation types: Towards automatic review article generation. *Advances in Classification Research Online*, 11(1):117–134.

Josh Nicholson, Milo Mordaunt, Patrice Lopez, Ashish Uppala, Dominic Rosati, Neves Rodrigues, Peter Grabitz, and Sean Rife. 2021. Scite: A smart citation index that displays the context of citations and classifies their intent using deep learning. *Quantitative Science Studies*, 2:1–38.

Son Bao Pham and Achim Hoffmann. 2003. A new approach for scientific citation classification using cue phrases. In *Australasian Joint Conference on Artificial Intelligence*, pages 759–771. Springer.

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

Hassan Raza, M. Faizan, Ahsan Hamza, Ahmed Mushtaq, and Naeem Akhtar. 2019. Scientific text sentiment analysis using machine learning techniques. *International Journal of Advanced Computer Science and Applications*, 10(12).

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

Ruan Visser and Marcel Dunaiski. 2022. Sentiment and intent classification of in-text citations using bert. EasyChair Preprint no. 7593.

Wenhao Yu, Mengxia Yu, Tong Zhao, and Meng Jiang. 2020. Identifying referential intention with heterogeneous contexts. pages 962–972.