Scientometrics (2023) 128:735–759
https://doi.org/10.1007/s11192-022-04567-4

Understanding the meanings of citations using sentiment, role, and citation function classifications

Indra Budi[1] · **Yaniasih Yaniasih**[1,2]

Received: 30 May 2022 / Accepted: 13 October 2022 / Published online: 14 November 2022

Abstract
Traditional citation analyses use quantitative methods only, even though there is meaning in the sentences containing citations within the text. This article analyzes three citation meanings: sentiment, role, and function. We compare citation meanings patterns between fields of science and propose an appropriate deep learning model to classify the three meanings automatically at once. The data comes from Indonesian journal articles covering five different areas of science: food, energy, health, computer, and social science. The sentences in the article text were classified manually and used as training data for an automatic classification model. Several classic models were compared with the proposed multi-output convolutional neural network model. The manual classification revealed similar patterns in citation meaning across the science fields: (1) not many authors exhibit polarity when citing, (2) citations are still rarely used, and (3) citations are used mostly for introductions and establishing relations instead of for comparisons with and utilizing previous research. The proposed model's automatic classification metric achieved a macro F1 score of 0.80 for citation sentiment, 0.84 for citation role, and 0.88 for citation function. The model can classify minority classes well concerning the unbalanced dataset. A machine model that can classify several citation meanings automatically is essential for analyzing big data of journal citations.

Keywords Citation meaning · Citation sentiment · Citation role · Citation function · Convolutional neural network · Multi-output model

Mathematics Subject Classification 68T07 Artificial neural networks and deep learning · 68T50 Natural language processing

JEL Classification C45 Neural networks and related topics · C55 Large data sets: Modeling and analysis

✉ Indra Budi
indra@cs.ui.ac.id

[1] Faculty of Computer Science, Universitas Indonesia, Kampus UI, Depok, Jawa Barat 16424, Indonesia

[2] National Research and Innovation Agency Republic of Indonesia, Jl. M.H. Thamrin No. 8, Jakarta Pusat 10340, Indonesia

Introduction

Citations should be classified according to their use within the text, not only based on the bibliography, as is currently mostly the case (Moravcsik & Murugesan, 1975; Swales, 2004). Citation analysis has been widely used for evaluating research performance (Aksnes et al., 2019; Lukman et al., 2018), rankings (Aksnes et al., 2012; Massucci & Docampo, 2019), studies on scientific developments (Murillo et al., 2021; Pallottino et al., 2018), and plagiarism detection (Gipp et al., 2013), among others. However, these analyses are still primarily based on references in the bibliography. This method has been criticized as being biased, subjective, inconsistent and non-standardized, widely misused, and invalid (Anninos, 2014; Belter, 2015; Molas-Gallart & Ràfols, 2018; Wallin, 2005). Understanding citations' textual contexts helps improve the accuracy of analyses.

Citations in the text can be examined via their intensity (frequency of citation), location (in the introduction, method, or result section), and textual context (Boyack et al., 2018; Lu et al., 2017; Nazir et al., 2020; Yaniasih & Budi, 2021a; Zhao & Strotmann, 2020). The context can reveal the author's intent when citing an article in their writing is often referred to as the function or purpose of the citation. There are many citation function categories, such as "introducing," "relating to," "using," and "comparing with" other literature (Lin, 2018; Teufel et al., 2006). In addition, the author's opinion on the article can be addressed through sentiment, i.e., "positive," "negative," or "neutral" polarity (Ikram & Afzal, 2019; Yousif et al., 2019a, 2019b). Furthermore, the role of the cited article can be identified, be it "data," "method," or "supplemental" (Zhao et al., 2019).

Figure 1 presents an example of a citation in the text and its meanings. The sentence in the figure reads, "*Bertin* et al. *analyses 45.000 articles from PLOS journals. Their research found that the citation distribution in the text varies by journal series…"*. This sentence does not indicate the author's polarity, so the sentiment is recorded as "neutral". The sentence provides information about the research's finding, so the role is "result." Based on the sentence, it can also be seen that the author's purpose in citing these articles is to "relating" with the cited article.

In recent years, the citation context has been evaluated utilizing various data, methods, and discussions. Most of the evaluated data were articles from journals written in English and published in developed countries. Most of the topics in the journal are related to computer science (computational linguistics, bioinformatics, neural information) (Bakhti et al., 2018b; Cohan et al., 2019; Ikram & Afzal, 2019; Mercier et al., 2018; Rachman et al., 2019; Su et al., 2019; Tuarob et al., 2019; Wang et al., 2019; Yousif et al., 2019a, 2019b; Zhao et al., 2019), health sciences and medicine (Kilicoglu et al., 2019; Xu et al., 2015; Yan et al., 2019), library and information science (Aljuaid et al., 2020; Taskin & Al, 2017), and some natural science topics. Few studies utilize multiple domains since the majority employ a single domain. Regarding approach, citation contexts have been analyzed using manual and rule-based methods (Dehdarirad & Yaghtin, 2022), traditional machine learning (Aljuaid et al., 2020; Amjad & Ihsan, 2020), and deep learning (Muppidi et al., 2020; Zhang et al., 2022). Analyses simultaneously of two meanings have also been performed, such as sentiment and function (Huang et al., 2021; Jha et al., 2017; Jia, 2018; Yousif et al., 2019a, 2019b), as well as functions and role (Zhao et al., 2019). To fully comprehend the relationship and significance of a citation, it is necessary to recognize its three meanings together (Moravcsik & Murugesan, 1975). However, no single approach and discussion of three citation meanings have yet been discovered.

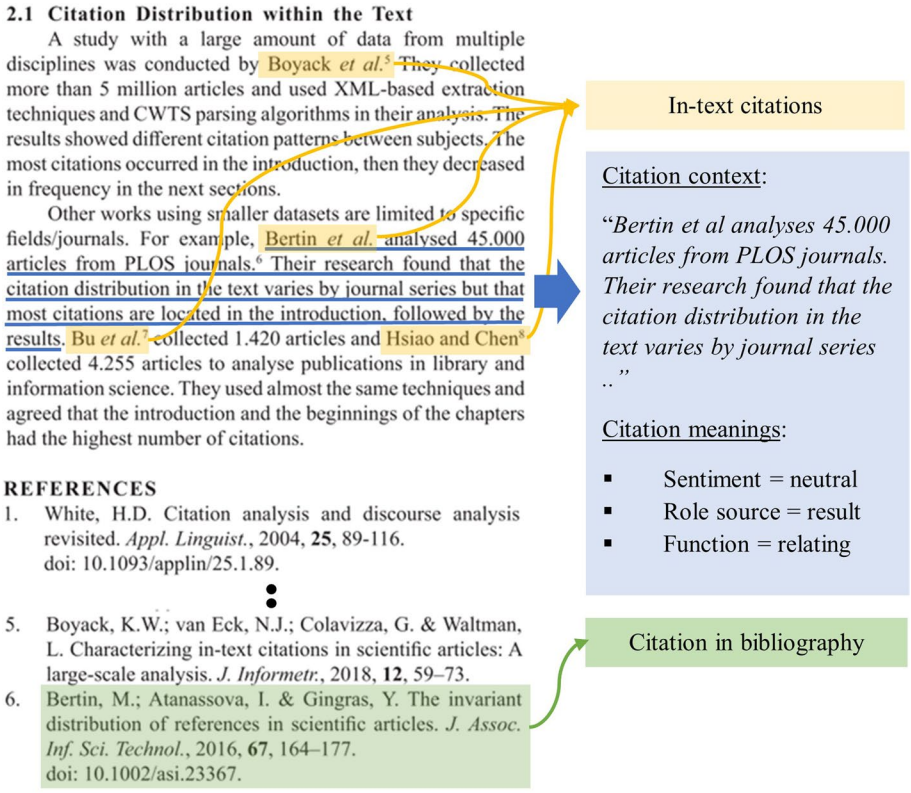


Fig. 1 Examples of citation context, in-text citation, and citation in a bibliography

This paper aims to analyze three citation contexts, i.e., sentiment, role, and function. The goal is to address the following research problems: (1) the pattern of three citation meanings in different scientific domains has not been extensively studied, and (2) there is currently no automatic model that can examine three citation meanings concurrently. The analysis is carried out in five fields of science: food, energy, health, computer, and social sciences. These five fields represent significant, yet substantially different, fields of science. From a technological perspective, this study proposes to perform simultaneous, automatic classification using a deep learning multi-output model and compare it to the existing state-of-the-art model (single-output approach). The multi-output model can provide more efficient and accurate classifications than the separate classification models. The novelty and contribution summary of this paper is presented in Table 1.

Literature review

Citation context

Citation analysis has been widely discussed and implemented in library and information science, computer science, and quantitative science studies. This analysis examines the

Table 1 The novelty and contribution summary

Citation context studies	The existing state of the art	Novelty and contribution
Data	Mostly used: International journals Published in developed countries Single or limited domains in one study	Complement the literature from different perspectives using: Indonesian journals Published in developing countries Multiple domains: computer, health, food, energy, and social science
Method	Human classification Computer-based classification: Rule-based Traditional machine learning Single-output deep learning Multi-output deep learning for two citation meanings	Propose a multi-output deep learning model for three citation meanings classification
Analysis	One or two citation meanings patterns	Analyze three citation meanings patterns at once

number, pattern, and network of citations in published documents. Citation analysis arose from the assumption that citations can provide information about the relationship between articles, the history of the idea development, and the discovery of specific research topics (De Bellis, 2010). The typical citation analysis so far is calculating citation numbers in the bibliography. This traditional method was considered less valid because it only measures the quantity, not the quality of citations (Shahid et al., 2015).

The in-text citation analysis has become a recommendation to improve the citation analysis method. There are three variables of in-text citation: intensity, location, and sentence context. The earliest reference of in-text citation research found that perfunctory citations were in the introduction section. Meanwhile, the essential citations were in the methodology, results, and discussion sections (Maricic et al., 1998). Another finding showed that citations in the methodology section were more relevant than those in the literature review section (Athar & Teufel, 2012).

The citation context variable analyzes the language meaning of the sentence containing citations. Moravcsik and Murugesan (1975) initially described the citation context analysis scheme. Based on the connection and the significance of a citation, they questioned what a citation meant. The relationship's meaning can be determined by (1) whether what is cited is conceptual or operational, and (2) whether it is a research base or an alternative (evolutionary or juxtapositional). Furthermore, (3) whether a citation is necessary or only for recognition (organic or perfunctory), and (4) whether it is accepted or rejected (confirmative or negational), determine the citation's value. This idea has become the main reference point for almost all literature on citation context classification. Point four evolves into citation sentiment analysis. Points two and three lead to an examination of the citation function. Sentiment analysis and citation functions are frequently investigated, discussed, and developed. Point one has become an analysis of citation role, but it hasn't been examined as much as citation sentiment and function analysis.

Citation sentiment

Sentiment analysis identifies and classifies opinions in text or image documents. This subject was placed in the early 2000s and experienced substantial growth after 2009 (Piryani et al., 2017). Product review sentiment, social media dialogues, news, and blogs are the most frequently evaluated areas. According to Yousif et al., (2019b) citation sentiment analysis on scientific articles was detected for the first time in 2011.

Citation sentiment analysis has emerged and is expanding. There are at least two key reasons why citation sentiment analysis is essential. The first is to improve bibliometric metrics by accounting for quality rather than quantity, minimizing citation bias, and offering authorship support based on scientific evidence. The second goal is to detect non-reproducible research, particularly in the biomedical field, where unfavorable attitudes might be an early indicator of research that is not reproducible, thereby saving research time and resources (Xu et al., 2015). However, Catalini et al. (2015) identified that even negative citations have a specific role in the scientific community. In some cases, negative citations can assist refine original discoveries and contribute to the overall development of a field.

Since its inception, manual and automatic classification using traditional machine learning has been done. Recent research was conducted by (Dehdarirad & Yaghtin, 2022), who classified citation sentiment manually in life science and biomedicine citations. Sentiment results were compared statistically between males and females, showing a scientific communication pattern. Several studies have demonstrated that the support vector machine

(SVM) model outperforms other machine learning methods to classify citation sentiment. Xu et al. (2015) classified 4182 sentences in clinical trial papers using SVM and obtained an F1 value of 0.71. Mercier et al. (2018) also got an F1 value of 0.71 using a combined multi-classifier between SVM and a perceptron on 2100 computer data sentences. SVM is also used by (Aljuaid et al., 2020) to classify 8736 sentences in the field of information science and got the highest F1 of 0.83. Another machine learning model used to classify a massive number of citation sentiments (762,355 datasets) is Naive Bayes. Still, the performance evaluation of the model used is not shown (Catalini et al., 2015).

The preprocessing method and manual feature selection substantially influence the results of classical machine learning models. Furthermore, citation sentiment analysis is challenging because the data is highly uneven, with the number of negative citations being far lower than in the other two classes (Ravi et al., 2018). This limitation promotes the use of deep learning approaches to solve current issues.

The deep learning model that is most frequently used for categorizing citation sentiment is convolutional neural networks (CNN). Kilicoglu et al. (2019) examined SVM, CNN, and BiLSTM rule-based models. The CNN model produced the most excellent results in the health area, with an F1 value of 0.72 on the 4182 datasets. Yousif et al., (2019a) acquired an F1 value of 0.88 on 5568 datasets in computer science utilizing a mixture of CNN and BiLSTM. Muppidi et al. (2020) used a combination of CNN, LSTM, and word2vec to perform sentiment classification on 7640 sentence data and obtained an F1 value of 0.85. Wang et al. (2019) achieved the best result of 0.93 utilizing CRF and CNN on 3500 computer science datasets. Table 2 shows some existing studies in citation sentiment analysis.

Citation function

Citation function analysis is well-studied. Most focus on category schemes and classification models. The function category scheme varies based on data attributes, classification goals, and use. Since the classified data was algorithm sentences, Tuarob et al. (2019) picked the function scheme which consisted of "utilized" and "not utilized." "Utilized"

Table 2 Existing citation sentiment literature

Literature	Dataset size	Domain/Topic	Method	Result (F1 score)
Yan et al. (2019)	12,000	Chemistry, physiology, medical sciences	SenticNet	0.67
Raza et al. (2020)	5161	NA	SVM	0.70
Xu et al.(2015)	4182	Clinical trial	SVM	0.71
Mercier et al. (2018)	2100	Computer science	SVM and a perceptron	0.71
Kilicoglu et al. (2019)	4182	Heath science	CNN	0.72
Ikram and Afzal (2019)	8736	Computer science	SVM	0.75
	4182	Bioinformatics		
Aljuaid et al. (2020)	8736	Information science	SVM	0.83
Muppidi et al. (2020)	7640	Autism	CNN and LSTM	0.85
Yousif et al. (2019a)	3568	Computer science	CNN and BiLSTM	0.88
Wang et al. (2019)	3500	Computer science	CRF and CNN	0.93

consisted of "use," "extend," and "not utilized" consisted of "mention" and "not algorithm." Cohan et al. (2019) picked three schema classes ("background/information," "technique comparisons," and "outcome comparisons") because they were necessary for exploring subjects, connected to the scientific article structure, and easy to execute using machine learning. Bakhti et al. (2018a) introduced a citation system with five functions: "useful," "contrast," "mathematical," "accurate," and "neutral." This generic categorization approach was relevant to many scientific disciplines and easily recognized by humans. Rachman et al. (2019) altered four citation functions ("problem," "other," "use data," "use model," "use tool") to construct a document-summarizing system. Yaniasih and Budi (2021b) used Indonesian journal types to quantify citation value for ranking science using five schemes ("background," "use," "extend," "compare," and "related"). This study adapts these schemes due to the data's comparability and the implementation's objective.

The automatic citation function categorization method extensively uses traditional machine learning and deep learning. However, most have been using a single output model approach, in which a model performs only one classification. While multiple citation meanings use the same data, it is possible to process them simultaneously using a multi-output model. The existing state of the art of citation function classification using both single output and multi-output models is presented in Table 3.

Table 3 compares single- and multi-output models. The best result for the single output model employing Naïve Bayes has an F1 score (0.78) (Taskin & Al, 2017), and using SVM has the most outstanding (0.90) (Tuarob et al., 2019). The multi-output model mostly used automatic feature deep learning and performed exceptionally well. Cohan et al. (2019) simultaneously achieved citation function and location classification using the structural scaffold features, Glove, and Elmo in multi-task learning bi-directional long-short term memory (BiLSTM). The result obtained an F1 score of 0.84. Another study was conducted by Su et al. (2019) for citation function and provenance using a convolutional neural network (CNN). Function accuracy was obtained at 0.69 and provenance at 0.79. Yousif et al., (2019a) got the experiment with the highest yield for citation sentiment and purpose classification. The model used combined CNN and BiLSTM, resulting in an F1 value of 0.88 for sentiment and 0.84 for citation purposes. Research by Zhao et al. (2019) used multi-task learning to classify roles and citation functions. The Recurrent Neural Network (RNN) with the BERT pre-trained model produced an F1 value of 78% better than some single-task models.

Citation role

Some studies did not distinguish between citation role and function meanings, combined them, or used them interchangeably as words with the same meaning. Kwan and Chan (2014) stated that the role of citation is identical to its function. Agarwal et al. (2010) designed a class schema of citation meanings and referred to them as role labels. Nevertheless, the label category encompassed a combination of roles (material/method) and functions (contemporary, contrast, evaluation, explanation, modality, and similarity). The phrase citation role by Jurgens et al. (2016) was of a higher level and can be separated into two meanings: centrality and citation function. The centrality of a reference reveals whether it is quoted because it plays a vital role or because the context is broader. This method resembles the citation role scheme by Bedi et al. (2022), which categorizes

Table 3 Existing citation function literature

No.	References	Dataset size (sentence)	Domain/topic	Function scheme	Method	Result (F1 score)
1	Bakhti et al. (2018b)	8700	Computational linguistics	Based on, supply, useful, acknowledgement, contrast, weakness, correct, hedges	CNN	0.65
2	Rachman et al. (2019)	1153	Computational linguistics	Problem, other, use data, use model, use tool	SVM	0.68
3	Su et al. (2019)	1432	Computer science	Weakness, compare and contrast, positive, and neutral	CNN	0.69
4	Taskin and Al (2017)	10,437	Library and information science	Literature, definition, using/explaining methods, comparison, mentioning pioneers, proof support, generate ideas for the future, criticizing, giving example, using data, data validation	Naïve bayes	0.78
5	Zhao et al. (2019)	2814	Computational linguistics, neural information, biomedical and life sciences	Use, produce, introduce, extend, compare, other	SciRes CLF	0.78
6	Cohan et al. (2019)	1941 and 11,020	Computational linguistics; computer science, and medicine	Background information, Method, Result comparisons	BiLSTM	0.84
7	Yousif et al., (2019a, 2019b)	3568 and 1768	Computer science	Neutral, criticizing, comparison, use, substantiating, basis, neutral; and neutral, idea, basis, related, compare	RCNN	0.88
8	Tuarob et al. (2019)	8796	Algorithm	Utilized: use, extend and not utilized: mention, not algorithm	SVM with content and context features	0.95

citations as baseline or non-baseline. When a source is cited, it belongs to the baseline class since it serves as the basis or comparison for the study.

Different from the research above, the term citation role in this article relates to the category of citation context by answering the issue of whether the meaning of the cited article is conceptual or operational (Moravcsik & Murugesan, 1975). Numerous studies have produced several classification schemes based on this principle. Considering the nature of computer journal articles, Guo et al. (2014) modified the idea and then divided the operational class into a method, dataset, and performance evaluation. They employed its scheme to classify 2156 sentences and yielded an F1 score of 0.53 using Random forest. This study advances a scheme from Zhao et al. (2019) that classified citation roles as data, tool, code, algorithm, document, website, paper, license, and media. These nine fine-grained classes are then aggregated into three more general categories: materials (data), techniques (tools, code, algorithms), and supplements (documents, websites, papers, licenses, and media). This category pertains to writing styles, particularly in computer science and engineering, where identifying tasks, techniques, and materials is crucial when attributing sources (Augenstein et al., 2017; Luan et al., 2017). Zhao et al. (2019) proposed a multi-tasking model called SciResCLF and obtained an F1 score of 0.78. Since the data in the study is not limited to the fields of computers or engineering, the citation scheme comprises data (material), method, result, and supplement.

The above review of the previous research identified some shortcomings of citation context analysis. The first drawback is that the data used are still limited in number, the scope of the domain, and language. Most of the research used citation data from computer science journals and a small part from health and medicine. Even though fields of science strongly influence citation characteristics (Levitt & Thelwall, 2008, 2009). If there is only scientific evidence from one or two specific domains, it will give a significant gap in the development of in-text citation analysis. In addition, almost all data sets were in English journals. This study attempts to fill the data gaps mentioned above. The data used in this study are citations in Indonesian journal articles of five science fields, namely food, energy, health, computer, and social. The second shortcoming of in-text citation analysis literature is that most existing research performs a manual or separate automatic classification of citation contexts. Few studies classify two citation meanings simultaneously. To the best of our knowledge, this paper is the first in library and information science and computer science to analyze three citation meanings: sentiment, role, and function at once.

Methods

The study consisted of four phases: data collection, manual classification, selection of automatic classification models, and model performance evaluation. Following is a detailed description of the process at each stage. Figure 2 illustrates all phases of the process.

Collection of datasets

The data analyzed were sentences containing citations in Indonesian scientific journal articles published in 2019. The journals came from five disciplines: food, energy, health, social, and computer science. They were processed using the Grobit parsing tool (Lopez, 2009), which converts PDF documents into lists of sentences ready to be classified. A total

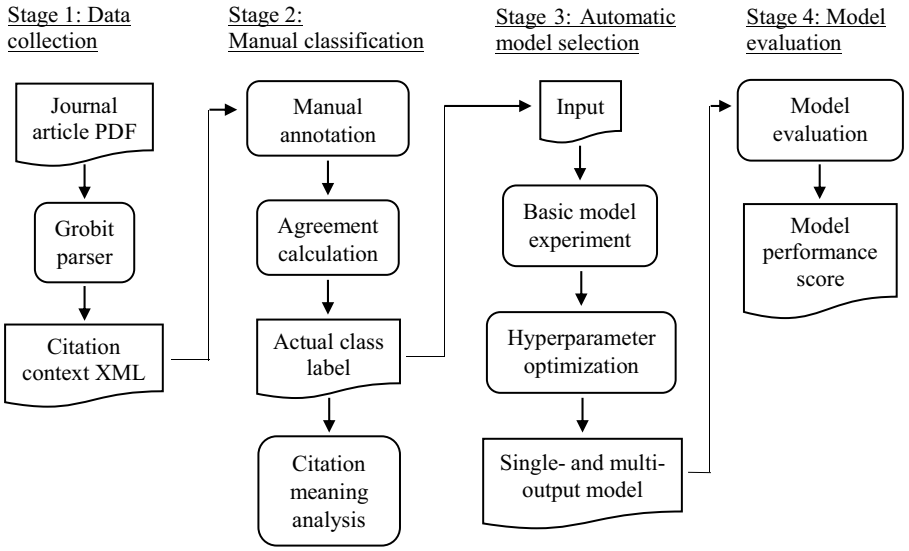


Fig. 2 The sequence of research phases

of 852 articles were processed, consisting of 9173 sentences. The statistics for the dataset are presented in Table 4.

The number of journals and articles analyzed was limited because the data set only included journals in the SINTA 1 and 2 categories. SINTA is a journal indexer that evaluates the quality of Indonesian journals (https://sinta.kemdikbud.go.id/). Despite SINTA's selection, there were still journals whose writing structure and format did not meet scientific writing standards and thus could not be processed further.

Classifying citation contexts employs smaller data sets than studying citation frequency and location. Citation context datasets require significant preprocessing, such as manual annotation, which may reduce the number of datasets due to label inequality or limited processing resources. The majority of previous studies examined less than 10,000 sentences. Raza et al. (2020) classified 5161 and 4989 sentences; Ikram and Afzal (2019) classified 8736 and 4182; Kilicoglu et al. (2019) classified 4182. Only Yan et al. (2019) used over 12,000 sentences. Perier-Camby et al. (2019) employed 3000 phrases for function classification. As for the classification of two citation meanings, Zhao et al. (2019) used 2814 phrases for roles and functions, while Yousif et al., (2019a) utilized 3568 and 1768 for sentiment and function, Su et al. (2019) classified 1432 and 1492 for source and function, and Cohan et al. (2019) used 1941 and 11,020 for location and function. Previous tables (1 and 2) present variations in the amount of data in the citation context analysis. Based on this circumstance, although the number of data sets in this study is limited, it is comparable and very substantial compared to most previous studies.

Manual classification

Big data is often involved in citation analyses, meaning that manually classifying this number of citations is impossible. Hence, they must be classified automatically via computer.

Table 4 Statistics for the dataset

Discipline	Journal	Number of articles	Number of sentences
Food	Journal of Nutrition and Food	18	595
	Journal of Food Technology and Industry	16	460
	Journal of Food Technology Applications	15	459
	Agricultural Postharvest Research Journal	15	224
	Journal of Agricultural Technology & Industry	15	328
	Advances in Food Science, Sustainable Agriculture and Agroindustrial Engineering	21	266
	Agricultural Products Industry News	13	331
Energy	Journal of Nuclear Energy Development	20	154
	Journal of Electricity and Renewable Energy	16	121
	Journal of Manufacturing Energy Engineering	21	79
	Scientific Journal of Energy and Electricity	27	92
	Journal of Natural Materials Engineering and Sustainable Energy	19	60
	Journal of Minerals, Energy and the Environment	24	119
	ELKOMIKA: Journal of Electrical Energy, Telecommunications, & Electronics Engineering	43	591
Health	Indonesian Journal of Health Promotion	27	67
	Journal of Health Ecology	38	365
	Journal of Vocational Health	42	271
	Health Research Bulletin	60	1028
	Indonesian Journal of Environmental Health	37	328
Social	Journal of Marine and Fisheries Socio-Economic Policy	53	226
	Journal of Marine and Fisheries Socio-Economic	54	408
	Marina Scientific Newsletter	53	282
	Journal of Social Psychology	51	415
	Sosio Koncepsia: Social Welfare Development and Research Journal	58	246

Table 4 (continued)

Discipline	Journal	Number of articles	Number of sentences
Computer	Journal of Computer Engineering System and Science	13	214
	Journal of Information Technology and Computer Engineering	6	76
	Journal of Computer Technology and Systems	18	583
	Journal of Information Technology and Computer Science	45	596
	Informatika Mulawarman: Scientific Journal of Computer Science	14	189
Total		852	9173

Small data sets with human annotations are needed as data training for computer algorithms to do automatic categorization.

In this stage, the collected dataset was first classified manually, i.e., class labels were assigned. Three people with similar educational backgrounds carried out the manual classification according to the scientific field. The labels for the sentiment were "positive" when the citation confirms the cited article, "negative" when the citation criticizes or rejects the cited article, and "neutral" when no polarity arises (Yousif et al., 2019b). The role labels consisted of "data," "method," "result," and "supplemental". The function labels included "introducing," "relating," "utilizing," "explaining," and "comparing." The function scheme improved on the previous research scheme (Yaniasih & Budi, 2021b), resulting in more balanced data.

The degree of agreement between the three annotators was measured using Fleiss Kappa values. The value for the sentiment was 0.69, the value for the role was 0.78, and the value for the function was 0.61, indicating substantial agreement between annotators (Landis & Koch, 1977). The data used were approved by at least two annotators and amounted to 8566 sentences. The result from manual classification, called actual class label, is presented in Fig. 3. An algorithm then learned the labeled data until it could correctly classify and predict the instances. The hope is that large amounts of actual data can be classified accurately.

Selection of automatic classification model

The model proposed for automatic classification involved a convolutional neural network (CNN). CNNs have been used extensively and successfully for processing images, text, and speech (Alom et al., 2019; Khamparia & Singh, 2019; Shrestha & Mahmood, 2019). Several studies have also shown that CNNs can classify citation meanings well (Bakhti et al. 2018b; Kilicoglu et al., 2019).

CNN consists of two main parts, namely feature extraction and classification. The feature extraction section consists of convolution and pooling (sub-sampling) layers. The

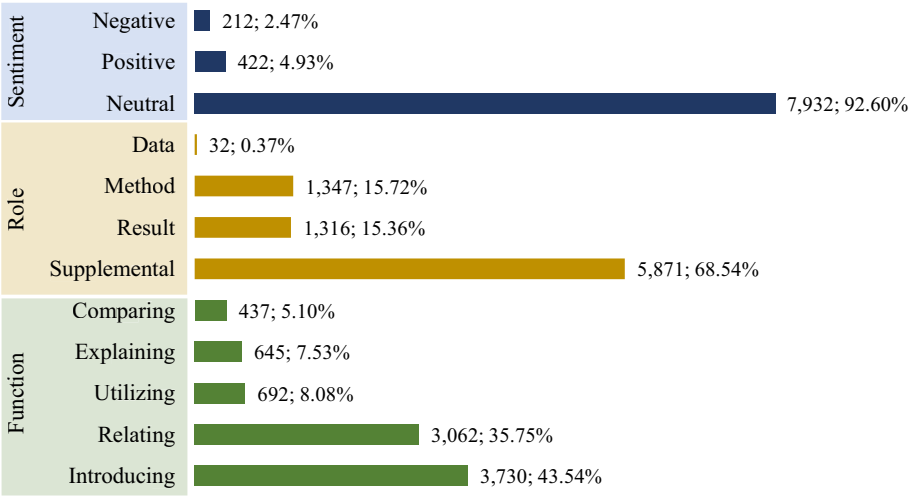


Fig. 3 Number per class label of citation context data set

convolution layer extracts data from a specific input part (in this study, the input is sentences). Each section's information is then mapped as features. Features are transmitted and passed on to the subsequent convolution layers, and a subsampling layer is utilized to obtain a more accurate representation of the features. The feature extraction layer's output becomes the classification layer's input. The classification layer is a fully connected network that uses multiple parameters to determine the score for each class. The network is trained to utilize gradient descent and backpropagation. The calculation uses a soft-max layer in which the class is determined by the highest score from each input (Alom et al., 2019; Khamparia & Singh, 2019; Shrestha & Mahmood, 2019). The fundamental structure of CNN is depicted in Fig. 4.

The model selection stage consisted of compiling the basic model, optimizing the hyperparameters, and evaluating the optimized model. The basic model consisted of the input, embedding, CNN, max-pooling, flattening, dense, dropout, and output layers. The input was a citation context sentence with three labels: sentiment, role, and function. In the embedding process, each word in the sentence was then represented as a numeric vector. Before classification, the word embedding was convoluted, and its dimensions were reduced.

The basic model had several hyperparameters that needed to be optimized to increase the model performance (Wu et al., 2019; Yang & Shami, 2020). The optimized hyperparameters included embedding, filter, kernel, dense unit, dropout rate, learning rate, and batch size. Optuna software was utilized for the optimization process because it can be used for both single-output and multi-output models, produces good performance outcomes, and provides various supporting features (Akiba et al., 2019). The optimal model was determined by the value for each hyperparameter that yields the lowest validation loss value.

After optimization was carried out on the basic model, the single and multi-output CNN models were obtained. The hyperparameter values and the best optimization results for the single- and multi-output models are presented in Table 5, and the architectures of these models are shown in Fig. 5.

Model performance evaluation

The optimized model was then evaluated for its performance and compared with several methods used in previous studies to classify citation sentiment, role, and function. The baseline models used for comparison were Nave Bayes (NB), Random Forrest (RF),

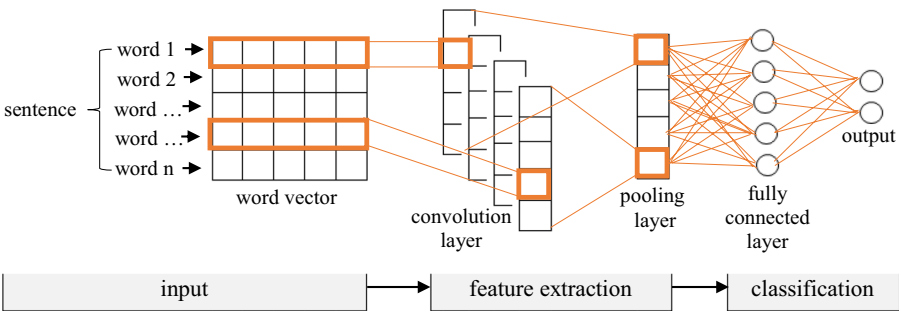
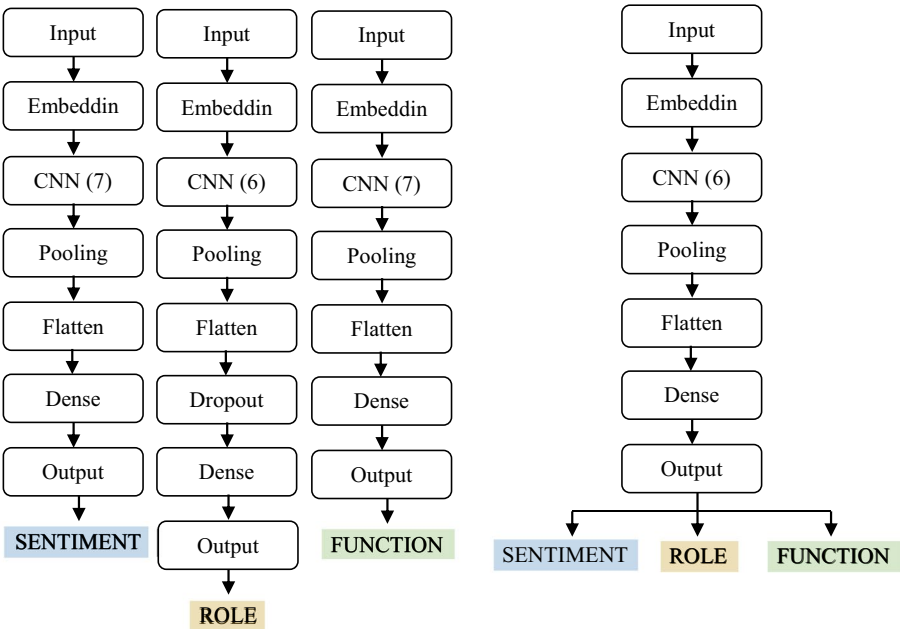


Fig. 4 CNN basic architecture for sentence classification

Table 5 Choices and best hyperparameters

Hyperparameter	Choices	Best			
		Single-ouput sentiment	Single- output role	Single-output function	Multi-output
Embedding size	32, 64, 128	128	64	32	32
CNN unit	1–10	7	6	7	2
Filter size	32, 64, 128	64	128	128	32
Kernel size	3, 4, 5	5	4	4	4
Dense unit	32, 64, 128	32	128	32	64
Dropout rate	0, 0.25, 0.5, 0.75	0	0.25	0	0
Learning rate	1, 0.1, 0.01, 0.001, 0.0001	0.001	0.001	0.001	0.001
Batch size	16, 32, 64	16	16	16	16
Smallest validation loss value		0.2175	0.5105	0.9458	2.3846



(a) Model 1: single-output (existing state of the art model)

(b) Model 2: multi-output (proposed model)

Fig. 5 Hyperparameter optimized models

Support Vector Machine (SVM), Long-short Term Memory (LSTM), and its Bidirectional model (Bi-LSTM). The training and machine validation process employed cross-validation. Classification ability was measured using the following metrics: accuracy, precision, recall, and macro F1 score (Lever et al., 2016). The mean macro takes the average across all classes regardless of class weight. In unbalanced data, the macro average will show

whether the model can detect minority classes well or not. The metric formula is depicted in Fig. 6 and Eqs. 1–4.

Results and discussion

Citation meanings in five fields of science

The citation sentiment patterns from the manual classification are almost identical across disciplines. On average, "neutral" has the highest percentage (92.60%), followed by "positive" (4.93%), then "negative" (2.47%). The "neutral" category accounts for 87–94% of citations across all disciplines, while "positive" ranks second with 4–9% of citations, except in computer science, where it ranks third with 1.21%. "Negative" classes contain few citations, with around 1% in food science, 2% in energy science, and 3% in both the health and social sciences. The citations per class are presented in Table 6.

These results of the five disciplines are generally the same as in previous studies, where most polarity classifications are "neutral", and the number of "negative" citations is always the smallest (Raza et al., 2020). A percentage of "negative" citations below 10% was also found by Xu et al. (2015) in a clinical trial paper, Jia (2018) using biomedical data, Catalini et al. (2015) specific in an immunology journal, and Huang et al. (2021) using a biological dataset. However, the number of "negative" citations in these Indonesian journals is lower than that found in the computer science field by Jha et al. (2017), where the percentage of "negative" sentiments reached 12%. In addition, a study by Yan et al. (2020) found that 15% of the citations were negative in the biomedical field.

The percentage of "negative" sentiments in scientific articles is low, presumably because researchers do not want to show their polarity to avoid confrontations with peers directly. Linguistically, sentences in scientific papers are official, so it is not easy to find sentences with polarity, unlike in product reviews and social media, where the language is more relaxed and expresses the authors' feelings (Hernandez-Alvarez et al., 2017; Jia, 2018).

		Predicted Class Label	
		Positive	Negative
Actual Class Label	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

$$Accuracy = \frac{\sum (TP + TN)}{\sum (TP + TN + FP + FN)} \quad (1)$$

$$Precision = \frac{\sum TP}{\sum (TP + FP)} \quad (2)$$

$$Recall = \frac{\sum TP}{\sum (TP + FN)} \quad (3)$$

$$F1\ score = \frac{\sum 2TP}{\sum (2TP + FP + FN)} \quad (3)$$

Fig. 6 Formula for calculating model performance metrics

Table 6 Citations per class

Classification	Class	Food (%)	Health (%)	Social (%)	Energy (%)	Computer (%)
Sentiment	Neutral	92.80	87.26	88.08	93.79	94.99
	Positive	5.66	9.58	8.62	4.20	1.21
	Negative	1.53	3.16	3.30	2.02	3.80
Role	Supplemental	67.18	70.78	73.24	63.98	63.67
	Result	15.27	24.36	15.66	9.74	10.32
	Method	17.55	4.18	10.91	25.86	25.47
	Data	0.00	0.73	0.19	0.42	0.54
Function	Introducing	32.59	45.50	41.09	46.35	46.41
	Relating	40.81	27.90	33.48	34.59	40.19
	Utilizing	10.68	4.28	9.00	10.92	6.16
	Explaining	9.52	14.00	12.94	6.63	3.62
	Comparing	6.31	8.31	3.42	1.43	3.62

Role assignments across the five disciplines are similar: the class with the highest number of citations is "supplemental", whereas the lowest number of citations is "data." The "supplemental" class contains 63–73% of the citations, while the "data" class contains a mere 1% across all five disciplines. Somewhat balanced percentages are seen for the "result" and "method" classes. The food, computer, and energy sciences have more "method" citations at 17.55%, 25.47%, and 25.86%, respectively, whereas the health and social sciences have more "result" citations at 24.36% and 15.66%, respectively. These findings differ from the research conducted by Zhao et al. (2019), with citations in the computer and health sciences in the "data" class at 31%, while those in the "supplement" class were at 30%.

A role can be related to citation location. For example, citations in the methods section usually cite "method" or "data," but citations in the results chapter cite the "results" of articles for comparison. Previous research that analyzed food journals showed that the percentage of citations in the methods section was around 9–16%, whereas the results and discussion section contained 43–54% of the citations (Yaniasih & Budi, 2021a). Manual classification of this research assigned approximately 17% of the citations to the "method" class and 15% to the "result" class. The "method" percentage is not much different from that in the previous research, but the "result" percentage is much lower. This difference is probably due to the authors using supplements to explain material in the results and discussion section. The low number of citations in the "data" class shows that citing data is still rarely done in various fields of science, as also reported by Liu (2015). However, with the increasing amount of data available in the digital era, citing data has become very important (Silvello, 2018).

The pattern for citation function is the same for the highest and second-highest percentages in the health, social, computer, and energy disciplines. The highest percentages for "introducing" in these three disciplines are recorded at 45.50%, 41.09%, 46.41%, and 46.35%, respectively. The second-highest percentages are assigned to "relating," with percentages in the range of 27–40%. The third-largest health and social sciences class is "explaining" at 12–14%. As for energy and computer science, the "utilizing" class holds the third position at 6–10%. "Utilizing" occupies the fourth position in the social sciences (9.00%) and the fifth (lowest) position in the health sciences (4.28%). The "comparing"

class occupies the lowest position in the social, energy, computer, and food sciences. In food science, the order is "relating" (40.81%), "introducing" (32.59%), "utilizing" (10.68%), and "explaining" (9.52%).

The citation function pattern in the five disciplines shows that citations function more as an "introducing" and "relating" with the cited literature. Based on the typology of citation quality (Moravcsik & Murugesan, 1975), the dominant number of "introducing" reveals that many citations are perfunctory and not entirely needed in the research process of citing articles. This pattern is also found in several studies where the number of perfunctory citations was quite large (Jurgens et al., 2018; Shu et al., 2019). The function of "relating" is higher than "using," indicating that the articles' connection is more conceptual or theoretical than operational. This finding is reinforced by the number of "comparing", which is lower than "explaining". "Comparing" functioned to compare the value of the results, while "explaining" worked to discuss the results using concepts or theories. Another research included functions other than "introducing" into essential citations (Lin, 2018). Consequently, the total percentage of essential citations was higher than perfunctory citations. However, functions other than "introducing" should not be given equal value because they have different levels of importance.

Models performance evaluations

Previously, traditional machine learning and single output deep learning models were frequently used in citation context research. On the other hand, this paper proposes a novel multi-output approach for classifying three citation meanings. The findings of this study indicate that the multi-output model employing CNN architecture performs better than the classic models. Table 7 compares the proposed model's performance to the existing state-of-the-art models as the baseline comparison.

All models achieved between 0.90 and 0.97 accuracies when classifying sentiment data. The accuracy value describes the classification accuracy of the model. Precision and recall are essential since sentiment data is imbalanced between classes. Precision describes the accuracy between requested and projected results. Recall value is the system's retrieval success rate. Precision and recall for the NB, LSTM (single- and multi-output), and BiLSTM (single-output) models were poor (<0.60), indicating the model might not classify reliably. The LR, SVM, and multi-output Bi-LSTM models had good precision but low recall (<0.60), meaning they might classify well but did not locate much accurate information. Single- and multi-output CNNs were accurate and reliable. These two models got 0.85 and 0.80 F1 values. Single-output models obtained higher recall and F1 values. However, the multi-output model was more precise.

Single- and multi-output CNN models had higher accuracy, precision, recall, and F1 for role classification. The F1 values for the CNN multi-output model and CNN single-output model were 0.84 and 0.81, respectively. Unlike the sentiment classification, all models' evaluation measure values were fairly good (>0.60). The classic machine learning models performed well, particularly LR, SVM, and RF, achieving accuracy, precision, recall, and F1 values of 0.83. With the maximum accuracy and precision values, the single-output Bi-LSTM model also worked well. However, because its recall value was low, the F1 value was lower than that of the CNN models.

Deep learning models did better than classical machine learning at function categorization. All traditional machine learning models scored F1 below 0.60. All deep learning

Table 7 Performance comparation between classic models and proposed model

Classification	Method	A	P	R	F
Sentiment	Single output NB + n-gram vector	0.90	0.49	0.45	0.47
	Single output LR + n-gram vector	0.94	0.71	0.59	0.63
	Single output SVM + n-gram TF-IDF	0.93	0.72	0.58	0.63
	Single output RF + n-gram vector	0.94	**0.89**	0.51	0.59
	Single output LSTM	0.95	0.52	0.52	0.52
	Single output BiLSTM	0.94	0.85	0.48	0.50
	Single output CNN	0.91	0.86	**0.87**	**0.85**
	Multi-output LSTM	0.94	0.52	0.46	0.48
	Multi-output BiLSTM	0.94	0.52	0.48	0.50
	Multi-output CNN (*proposed model*)	**0.97**	**0.89**	0.75	0.80
Role source	Single output NB + n-gram vector	0.77	0.78	0.77	0.77
	Single output LR + n-gram vector	0.83	0.83	**0.83**	0.83
	Single output SVM + n-gram TF-IDF	0.83	0.83	**0.83**	0.83
	Single output RF + n-gram vector	0.83	0.83	**0.83**	0.83
	Single output LSTM	0.96	0.71	0.70	0.71
	Single output BiLSTM	**0.97**	**0.97**	0.73	0.75
	Single output CNN	0.95	0.92	0.77	0.81
	Multi-output LSTM	0.88	0.64	0.60	0.60
	Multi-output BiLSTM	0.92	0.66	0.68	0.67
	Multi-output CNN (*proposed model*)	0.96	0.94	0.77	**0.84**
Function	Single output NB + n-gram vector	0.60	0.52	0.48	0.50
	Single output LR + n-gram vector	0.64	0.61	0.55	0.58
	Single output SVM + n-gram TF-IDF	0.63	0.59	0.55	0.57
	Single output RF + n-gram vector	0.65	0.72	0.51	0.54
	Single output LSTM	0.90	0.83	0.83	0.83
	Single output BiLSTM	0.93	0.89	0.82	0.85
	Single output CNN	0.88	0.83	0.78	0.80
	Multi-output LSTM	0.86	0.87	0.80	0.82
	Multi-output BiLSTM	**0.92**	**0.92**	0.87	0.87
	Multi-output CNN (*proposed model*)	0.91	0.87	**0.89**	**0.88**

A accuracy, *P* precision, *R* recall, *F* F1 score

Bold indicates the highest score

models had an F1 value over 0.80. The proposed model, multi-output CNN, got the highest F1 score of 0.88.

The multi-output CNN model is superior for role and function classification, while the single-output CNN model best does sentiment classification. Multi-output models are increasingly being used because there are many instances in which a single input is to complete several tasks simultaneously (Xu et al., 2020). One of the goals of any multi-output model is efficiency. The experimental results show that the multi-output model is more efficient in terms of training time, taking about 10% of the single output model's time for completion.

A more in-depth analysis was conducted on the multi-output model, which had the best performance. The investigation centered on classification performance per class.

Because the categories were unbalanced, attention was given to the model's ability to classify minority classes. For example, "positive" citations should receive a higher weight than "neutral" citations, whereas "negative" citations should be given lower weight than "neutral" citations in citation analyses (Abu-jbara et al., 2013; Kazi & Patwardhan, 2016). For role, "data" and "method" citations should receive greater weight than "supplemental" citations. The multi-output model successfully classified all classes well, including the minority classes (> 50%). Categories with large amounts of data, such as the "neutral" and "supplemental" categories, and all categories in the function classification, obtained F1 scores above 80%. The smallest class got the recall value, the lowest F1 score was negative,

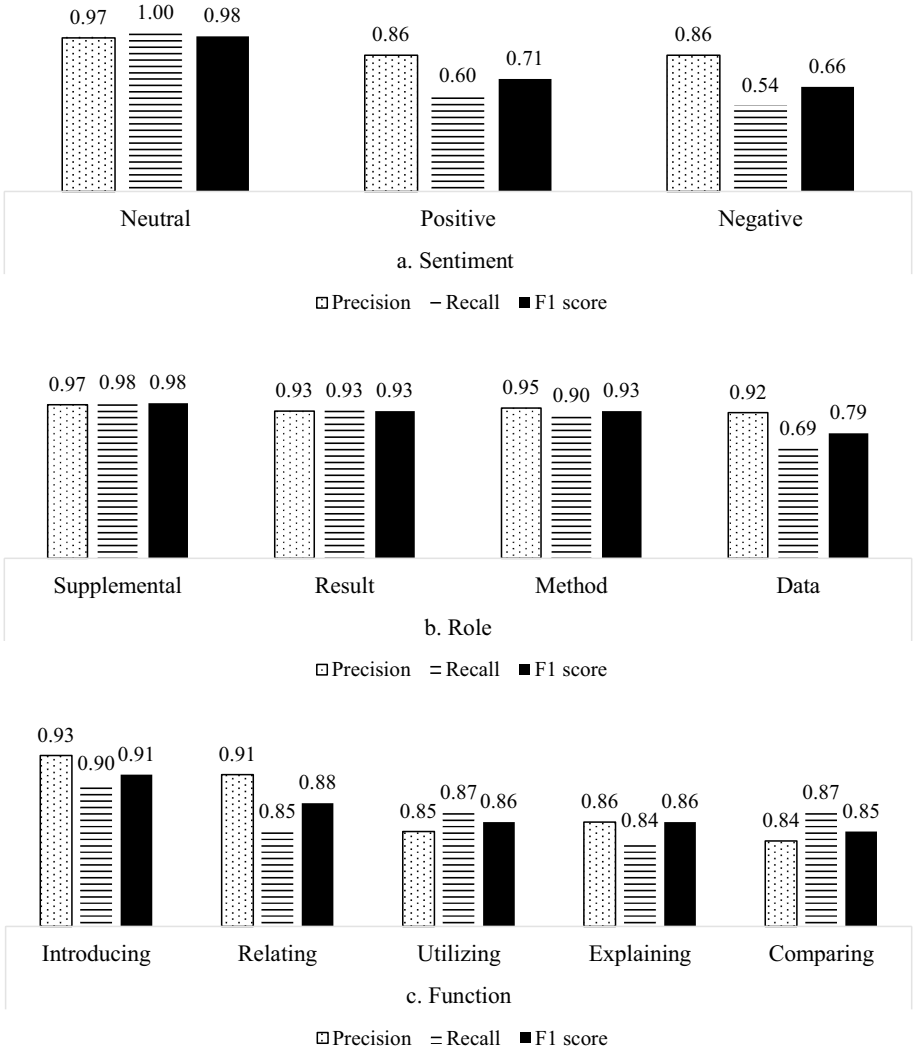


Fig. 7 The evaluation metrics for the classes

and the data both got an F1 score of 0.66. Figure 7 shows the evaluation metric values for the classes.

In this study, three citation meanings were analyzed using manual classification, followed by constructing a multi-output model for automatic classification. The findings included details on citation patterns in five academic fields and successfully proposed a deep learning model that performs better than the classic model. However, there are still some shortcomings with this study. First off, the scope and quantity of data are restricted to Indonesian journals. This coverage has disadvantages because citation patterns are influenced by culture, scientific fields, and other factors. Still, it also has benefits because, up to now, data from international journals from affluent nations have dominated citation research. This study's results can enhance non-developed country citation portraits. The second is the class category of role sources referring to Zhao et al. (2019) without doing a preliminary study on its appropriateness with the Indonesian journal writing style. Actually, it's possible that the proper approach for Indonesian journals differs from the one discussed. The annotators' agreement was moderate because many sentences don't fit the existing class structure. Increasing data's scope and altering the role source category schema can fix the problems. Meanwhile, procedures and models that have been developed can be employed again because the outcomes have been successful.

Conclusion

The analysis of sentences containing citations can identify the author's purpose in citing these articles, the author's opinions concerning the cited articles, and the roles of the articles being cited. To date, the analyses of these three meanings of citations have been carried out separately. It is essential to that a simultaneous analysis be carried out to improve the quality and efficiency of the citation analysis method.

The manual classification of the sentiment, role, and function of citations provided information on the meanings of the citations in several fields of science. Citation sentiment had the same pattern in the five disciplines analyzed: most of the citations were "neutral," only a few were "positive," and very few were "negative." Role classification followed the same pattern, where most of the citations were "supplemental," and very few were for "data." Citation function varied between disciplines, but it can be concluded that most fall under "introducing" and "relating," while few fall under "utilizing" and "comparing." The analysis above reveals that it is still rare for authors to show polarity in citing articles, data citation is rare, and authors use citations for introducing and relating more than for comparing and utilizing.

Automatic classification of three meanings can be done using traditional machine learning, single-output and multi-output deep learning models. The evaluation results show that the multi-output model utilizing CNN architecture outperforms the classic models for role and function classification but turns in slightly lower performance for sentiment classification. The capability of the multi-output CNN model is also quite good for minority classes, so it can be concluded that the model has good performance.

Author contributions Conceptualization: IB, YY; Methodology: YY; Analysis: IB, YY; Writing—original draft preparation: YY; Writing—review and editing: IB, YY; Funding acquisition: IB.

Funding This study was supported by a research grant from Universitas Indonesia (Hibah Publikasi Terindeks Internasional (PUTI) Q1 Tahun Anggaran 2022-2023 No: NKB-394/UN2.RST/HKP.05.00/2022).

References

Abu-jbara, A., Ezra, J., & Radev, D. (2013). Purpose and polarity of citation : Towards NLP-based bibliometrics. *Proceedings of NAACL-HLT*, *June*, 596–606.

Agarwal, S., Choubey, L., & Yu, H. (2010). Automatically classifying the role of citations in biomedical articles. *AMIA Annu Symp Proc.*, 11–15.

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna : A next-generation hyperparameter optimization framework. *Applied data science track paper, SIGKDD conference*, 2623–2631.

Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. *SAGE Open, 9*(1), 1–17. https://doi.org/10.1177/2158244019829575

Aksnes, D. W., Schneider, J. W., & Gunnarsson, M. (2012). Ranking national research systems by citation indicators. A comparative analysis using whole and fractionalised counting methods. *Journal of Informetrics, 6*(1), 36–43. https://doi.org/10.1016/j.joi.2011.08.002

Aljuaid, H., Iftikhar, R., Ahmad, S., Asif, M., & Afzal, M. T. (2020). Important citation identification using sentiment analysis of in-text important citation. *Telematics and Informatics*. https://doi.org/10.1016/j.tele.2020.101492

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Hasan, M., Van Essen, B. C., Awwal, A. A. S., & Asari, V. K. (2019). A state-of-the-art survey on deep learning theory and architectures. *Electronics (switzerland), 8*(3), 1–67. https://doi.org/10.3390/electronics8030292

Amjad, Z., & Ihsan, I. (2020). VerbNet based citation sentiment class assignment using machine learning. *International Journal of Advanced Computer Science and Applications, 11*(9), 621–627. https://doi.org/10.14569/IJACSA.2020.0110973

Anninos, L. N. (2014). Research performance evaluation: Some critical thoughts on standard bibliometric indicators. *Studies in Higher Education, 39*(9), 1542–1561. https://doi.org/10.1080/03075079.2013.801429

Athar, A., & Teufel, S. (2012). Context-enhanced citation sentiment detection. *NAACL HLT 2012—2012 conference of the northamerican chapter of the association for computational linguistics: human language technologies, proceedings of the conference*, 597–601.

Augenstein, I., Das, M., Riedel, S., Vikraman, L., & Mccallum, A. (2017). SemEval 2017 Task 10 : ScienceIE—Extracting keyphrases and relations from scientific publications. *546 Proceedings of the 11th international workshop on semantic evaluations (SemEval-2017)*, 546–555.

Bakhti, K., Niu, Z., & Nyamawe, A. S. (2018a). A new scheme for citation classification based on convolutional neural networks. *Proceedings of the international conference on software engineering and knowledge engineering, SEKE*, 131–142. https://doi.org/10.18293/seke2018-141

Bakhti, K., Niu, Z., & Yousif, A. (2018b). Citation function classi fi cation based on ontologies and convolutional neural networks. In L. Uden, D. Liberona, & J. Ristvej (Eds.), *Learning technology for education challenges. LTEC 2018. Communications in computer and information science* (Vol. 3, pp. 105–115). Springer. https://doi.org/10.1007/978-3-319-95522-3

Bedi, M., Pandey, T., Bhatia, S., & Chakraborty, T. (2022). Why did you not compare with that ? *Advances in information retrieval: 44th European conference on IR research, ECIR 2022*, 51–64.

Belter, C. W. (2015). Bibliometric indicators: Opportunities and limits. *Journal of the Medical Library Association, 103*(4), 219–221. https://doi.org/10.3163/1536-5050.103.4.014

Boyack, K. W., Jan, N., Eck, V., Colavizza, G., & Waltman, L. (2018). Characterizing in-text citations in scientific articles: A large-scale analysis. *Journal of Informetrics, 12*(1), 59–73. https://doi.org/10.1016/j.joi.2017.11.005

Catalini, C., Lacetera, N., & Oettl, A. (2015). The incidence and role of negative citations in science. *Proceedings of the National Academy of Sciences of the United States of America, 112*(45), 13823–13826. https://doi.org/10.1073/pnas.1502280112

Cohan, A., Ammar, W., Zuylen, M. Van, & Cady, F. (2019). Structural scaffolds for citation intent classification in scientific publications. *Proceedings OfNAACL-HLT 2019*, 3586–3596.

De Bellis, N. (2010). *Bibliometrics and citation analysis: From the science citation index to cybermetrics (Vol. 23, Issue 3)*. The Scarecrow Press. https://doi.org/10.1087/20100312

Dehdarirad, T., & Yaghtin, M. (2022). Gender differences in citation sentiment: A case study in life sciences and biomedicine. *Journal of Information Science*. https://doi.org/10.1177/01655515221074327

Gipp, B., Meuschke, N., Breitinger, C., Lipinski, M., & Nürnberger, A. (2013). Demonstration of citation pattern analysis for plagiarism detection. *SIGIR, 2013*, 1119–1120. https://doi.org/10.1145/2484028.2484214

Guo, C., Yu, Y., Sanjari, A., & Liu, X. (2014). Citation role labeling via local, pairwise, and global features. *77th ASIS&T annual meeting*, 1–10.

Hernandez-Alvarez, M., Soriano, J., & Martinez-barco, P. (2017). Citation function, polarity and influence classification. *Natural Language Engineering*. https://doi.org/10.1017/S1351324916000346

Huang, H., Zhu, D., & Wang, X. (2021). Evaluating scientific impact of publications: Combining citation polarity and purpose. *Scientometrics*. https://doi.org/10.1007/s11192-021-04183-8

Ikram, M. T., & Afzal, M. T. (2019). Aspect based citation sentiment analysis using linguistic patterns for better comprehension of scientific knowledge. *Scientometrics, 119*(1), 73–95. https://doi.org/10.1007/s11192-019-03028-9

Jha, R., Jbara, A. A., Qazvinian, V., & Radev, D. R. (2017). NLP-driven citation analysis for scientometrics. *Natural Language Engineering, 23*(1), 93–130. https://doi.org/10.1017/S1351324915000443

Jia, M. (2018). *Citation function and polarity classification in biomedical papers*. University ofWestern Ontario.

Jurgens, D., Hoover, R., & Mcfarland, D. (2018). Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics, 6*, 391–406.

Jurgens, D., Kumar, S., Hoover, R., McFarland, D., & Jurafsky, D. (2016). *Citation classification for behavioral analysis of a scientific field*. https://doi.org/10.48550/arXiv.1609.00435

Kazi, P. A. H., & Patwardhan, M. S. (2016). Context based citation summary of research articles: A step towards qualitative citation index. *IEEE international conference on computer communication and control, IC4 2015*. https://doi.org/10.1109/IC4.2015.7375701

Khamparia, A., & Singh, K. M. (2019). A systematic review on deep learning architectures and applications. *Expert Systems, 36*(3), 1–22. https://doi.org/10.1111/exsy.12400

Kilicoglu, H., Peng, Z., Tafreshi, S., Tran, T., Rosemblat, G., & Schneider, J. (2019). Confirm or refute?: A comparative study on citation sentiment classification in clinical research publications. *Journal of Biomedical Informatics, 91*(January), 103123. https://doi.org/10.1016/j.jbi.2019.103123

Kwan, S. B. C., & Chan, H. (2014). An investigation of source use in the results and the closing sections of empirical articles in information systems: In search of a functional-semantic citation typology for pedagogical purposes. *Journal of English for Academic Purposes, 14*, 29–47. https://doi.org/10.1016/j.jeap.2013.11.004

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174. https://doi.org/10.2307/2529310

Lever, J., Krzywinski, M., & Altman, N. (2016). Classification evaluation. *Nature Methods, 13*(8), 603–605.

Levitt, J. M., & Thelwall, M. (2008). Patterns of annual citation of highly cited articles and the prediction of their citation ranking: A comparison across subjects. *Scientometrics, 77*(1), 41–60. https://doi.org/10.1007/s11192-007-1946-y

Levitt, J. M., & Thelwall, M. (2009). The most highly cited library and information science articles: Interdisciplinarity, first authors and citation patterns. *Scientometrics, 78*(1), 45–67. https://doi.org/10.1007/s11192-007-1927-1

Lin, C. (2018). An analysis of citation functions in the humanities and social sciences research from the perspective of problematic citation analysis assumptions. *Scientometrics, 116*(2), 797–813. https://doi.org/10.1007/s11192-018-2770-2

Liu, X. (2015). Analyzing data citation practices using the data citation index nicolas. *Journal of the Association for Information Science and Technology*. https://doi.org/10.1002/asi

Lopez, P. (2009). GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), 5714 LNCS*, 473–474. https://doi.org/10.1007/978-3-642-04346-8_62

Lu, C., Ding, Y., & Zhang, C. (2017). Understanding the impact change of a highly cited article: A content-based citation analysis. *Scientometrics, 112*(2), 927–945. https://doi.org/10.1007/s11192-017-2398-7

Luan, Y., Ostendorf, M., & Hajishirzi, H. (2017). Scientific information extraction with semi-supervised neural tagging. *Proceedings of the 2017 conference on empirical methods in natural language processing*, *task 10*, 2641–2651.

Lukman, L., Dimyati, M., Rianto, Y., Subroto, I. M. I., Sutikno, T., Hidayat, D. S., Nadhiroh, I. M., Stiawan, D., Haviana, S. F. C., Heryanto, A., & Yuliansyah, H. (2018). Proposal of the S-score for measuring the performance of researchers, institutions, and journals in Indonesia. *Science Editing, 5*(2), 135–141. https://doi.org/10.6087/KCSE.138

Maricic, S., Spaventi, J., Pavicic, L., & Pifat-mrzljak, G. (1998). Citation context versus the frequency counts of citation history. *Journal of American Society for Information Science, 49*(6), 530–540. https://doi.org/10.1002/(SICI)1097-4571(19980501)49:6%3c530::AID-ASI5%3e3.0.CO;2-8

Massucci, F. A., & Docampo, D. (2019). Measuring the academic reputation through citation networks via PageRank. *Journal of Informetrics, 13*(1), 185–201. https://doi.org/10.1016/j.joi.2018.12.001

Mercier, D., Bhardwaj, A., Dengel, A., & Ahmed, S. (2018). SentiCite an approach for publication sentiment analysis. *ICAART 2018—Proceedings of the 10th international conference on agents and artificial intelligence*, *2*(Icaart), 422–429. https://doi.org/10.5220/0006587604220429

Molas-Gallart, J., & Ràfols, I. (2018). Why bibliometric indicators break down: Unstable parameters, incorrect models and irrelevant properties. *BiD*. https://doi.org/10.1344/BiD2018.40.23

Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science, 5*(1), 86–92.

Muppidi, S., Keerthi, S., & Kishore, B. (2020). An approach for bibliographic citation sentiment analysis using deep learning. *International Journal of Knowledge-Based and Intelligent Engineering Systems, 24*, 353–362. https://doi.org/10.3233/KES-200087

Murillo, J., Villegas, L. M., Ulloa-Murillo, L. M., & Rodríguez, A. R. (2021). Recent trends on omics and bioinformatics approaches to study SARS-CoV-2: A bibliometric analysis and mini-review. *Computers in Biology and Medicine, 128*(August 2020), 104162. https://doi.org/10.1016/j.compbiomed.2020.104162

Nazir, S., Asif, M., Ahmad, S., Bukhari, F., Afzal, M. T., & Aljuaid, H. (2020). Important citation identification by exploiting content and section-wise in-text citation count. *PLoS ONE, 15*(3), 1–19. https://doi.org/10.1371/journal.pone.0228885

Pallottino, F., Biocca, M., & Nardi, P. (2018). Science mapping approach to analyze the research evolution on precision agriculture: World, EU and Italian situation. *Precision Agriculture, 19*(6), 1011–1026. https://doi.org/10.1007/s11119-018-9569-2

Perier-Camby, J., Bertin, M., Atanassova, I., & Armetta, F. (2019). A preliminary study to compare deep learning with rule-based approaches for citation classification. *CEUR Workshop Proceedings, 2345*, 125–131.

Piryani, R., Madhavi, D., & Singh, V. K. (2017). Analytical mapping of opinion mining and sentiment analysis research during 2000–2015. *Information Processing and Management, 53*(1), 122–150. https://doi.org/10.1016/j.ipm.2016.07.001

Rachman, G. H., Khodra, M. L., & Widyantoro, D. H. (2019). Classification of citation sentence for filtering scientific references. *2019 4th international conference on information technology, information systems and electrical engineering, ICITISEE 2019*, 347–352. https://doi.org/10.1109/ICITISEE48480.2019.9003736

Ravi, K., Setlur, S., Ravi, V., & Govindaraju, V. (2018). Article citation sentiment analysis using deep learning. *Proceedings of 2018 IEEE 17th international conference on cognitive informatics and cognitive computing, ICCICC 2018*, 78–85. https://doi.org/10.1109/ICCI-CC.2018.8482054

Raza, H., Faizan, M., Akhtar, N., Abbas, A., & Naveed-Ul-Hassan. (2020). Scientific VS non-scientific citation annotational complexity analysis using machine learning classifiers. *International Journal of Advanced Computer Science and Applications, 11*(2), 210–213. https://doi.org/10.14569/ijacsa.2020.0110228

Shahid, A., Afzal, M. T., & Qadir, M. A. (2015). Lessons learned: The complexity of accurate identification of in-text citations. *International Arab Journal of Information Technology, 12*(5), 481–488.

Shrestha, A., & Mahmood, A. (2019). Review of deep learning algorithms and architectures. *IEEE Access, 7*, 53040–53065. https://doi.org/10.1109/ACCESS.2019.2912200

Shu, F., Qiu, J., Shu, S., & Larivière, V. (2019). Exploring the function of citations in ancient Chinese literature. *Proceedings of the Association for Information Science and Technology, 56*(1), 472–476. https://doi.org/10.1002/pra2.50

Silvello, G. (2018). Theory and practice of data citation. *Journal of the Association for Information Science and Technology, 69*(1), 6–20. https://doi.org/10.1002/asi.23917

Su, X., Prasad, A., Kan, M., & Sugiyama, K. (2019). Neural multi-task learning for citation function and provenance. *ACM/IEEE joint conference on digital libraries (JCDL) neural*, 394–395. https://doi.org/10.1109/JCDL.2019.00122

Swales, J. (2004). Citation analysis and discourse analysis revisited. *Applied Linguistics, 25*(1), 89–116+132. https://doi.org/10.1093/applin/25.1.89

Taskin, Z., & Al, U. (2017). A content-based citation analysis study based on text categorization. *Scientometrics, 114*(1), 335–357. https://doi.org/10.1007/s11192-017-2560-2

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). An annotation scheme for citation function. *Proceedings of the 7th SIGdial workshop on discourse and dialogue*, 80–87.

Tuarob, S., Kang, S. W., Wettayakorn, P., Pornprasit, C., Sachati, T., Hassan, S.-U., & Haddawy, P. (2019). Automatic classification of algorithm citation functions in scientific literature. *IEEE Transactions on Knowledge and Data Engineering*. https://doi.org/10.1109/TKDE.2019.2913376

Wallin, J. A. (2005). Bibliometric methods: Pitfalls and possibilities. *Basic and Clinical Pharmacology and Toxicology, 97*(5), 261–275. https://doi.org/10.1111/j.1742-7843.2005.pto_139.x

Wang, M., Leng, D., Ren, J., Zeng, Y., & Chen, G. (2019). Sentiment classification based on linguistic patterns in citation context. *Current Science, 117*(4), 606.

Wu, J., Hao, X. C., Xiong, Z. L., & Lei, H. (2019). Hyperparameter optimization for machine learning models based on Bayesian optimization. *Journal of Electronic Science and Technology, 17*(1), 26–40. https://doi.org/10.11989/JEST.1674-862X.80904120

Xu, D., Shi, Y., Tsang, I. W., Ong, Y. S., Gong, C., & Shen, X. (2020). Survey on multi-output learning. *IEEE Transactions on Neural Networks and Learning Systems, 31*(7), 2409–2429. https://doi.org/10.1109/TNNLS.2019.2945133

Xu, J., Zhang, Y., Wu, Y., Wang, J., Dong, X., & Xu, H. (2015). Citation sentiment analysis in clinical trial papers. *AMIA ... annual symposium proceedings. AMIA Symposium*, *2015*, 1334–1341.

Yan, E., Chen, Z., & Li, K. (2019). Authors' status and the perceived quality of their work: Measuring citation sentiment change in nobel articles. *Journal of the Association for Information Science and Technology, 00*, 1–11. https://doi.org/10.1002/asi.24237

Yan, E., Chen, Z., & Li, K. (2020). The relationship between journal citation impact and citation sentiment: A study of 32 million citances in PubMed Central. *Quantitative Science Studies, 1*(2), 664–674. https://doi.org/10.1162/qss

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing, 415*, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061

Yaniasih, Y., & Budi, I. (2021a). Analysis of in-text citation patterns in local journals for ranking scientific documents. *DESIDOC Journal of Library & Information Technology, 41*(2), 94–101.

Yaniasih, Y., & Budi, I. (2021b). Systematic design and evaluation of a citation function classification scheme in Indonesian journals. *Publications, 9*(27), 1–14.

Yousif, A., Niu, Z., Chambua, J., & Khan, Z. Y. (2019a). Multi-task learning model based on recurrent convolutional neural networks for citation sentiment and purpose classification. *Neurocomputing, 335*, 195–205. https://doi.org/10.1016/j.neucom.2019.01.021

Yousif, A., Niu, Z., Tarus, J. K., & Ahmad, A. (2019b). A survey on sentiment analysis of scientific citations. *Artificial Intelligence Review, 52*(3), 1805–1838. https://doi.org/10.1007/s10462-017-9597-8

Zhang, Y., Zhao, R., Wang, Y., Chen, H., & Mahmood, A. (2022). Towards employing native information in citation function classification. *Scientometrics*. https://doi.org/10.1007/s11192-021-04242-0

Zhao, D., & Strotmann, A. (2020). Deep and narrow impact: Introducing location filtered citation counting. *Scientometrics, 122*(1), 503–517. https://doi.org/10.1007/s11192-019-03280-z

Zhao, H., Luo, Z., Feng, C., Zheng, A., & Liu, X. (2019). A context-based framework for modeling the role and function of on-line resource citations in scientific literature. *EMNLP-IJCNLP 2019—2019 conference on empirical methods in natural language processing and 9th international joint conference on natural language processing, proceedings of the conference*, 5206–5215. https://doi.org/10.18653/v1/d19-1524