Hindawi
Complexity
Volume 2021, Article ID 5554874, 13 pages
https://doi.org/10.1155/2021/5554874

Research Article

Exploiting Contextual Word Embedding of Authorship and Title of Articles for Discovering Citation Intent Classification

Muhammad Roman,[1] **Abdul Shahid**,[1] **Muhammad Irfan Uddin**,[1] **Qiaozhi Hua**,[2] **and Shazia Maqsood**[1]

[1]*Institute of Computing, Kohat University of Science and Technology, Kohat, Pakistan*
[2]*Computer School, Hubei University of Arts and Science, Xiangyang 441000, China*

Correspondence should be addressed to Qiaozhi Hua; 11722@hbuas.edu.cn

Received 17 February 2021; Revised 13 March 2021; Accepted 23 March 2021; Published 5 April 2021

Academic Editor: Furqan Aziz

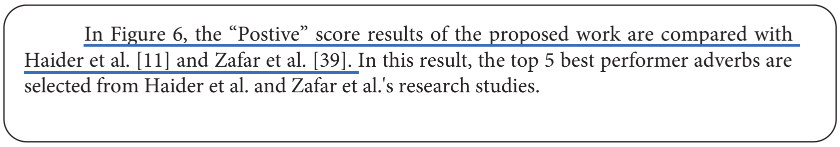
The number of scientific publications is growing exponentially. Research articles cite other work for various reasons and, therefore, have been studied extensively to associate documents. It is argued that not all references carry the same level of importance. It is essential to understand the reason for citation, called citation intent or function. Text information can contribute well if new natural language processing techniques are applied to capture the context of text data. In this paper, we have used contextualized word embedding to find the numerical representation of text features. We further investigated the performance of various machine-learning techniques on the numerical representation of text. The performance of each of the classifiers was evaluated on two state-of-the-art datasets containing the text features. In the case of the unbalanced dataset, we observed that the linear Support Vector Machine (SVM) achieved 86% accuracy for the "background" class, where the training was extensive. For the rest of the classes, including "motivation," "extension," and "future," the machine was trained on less than 100 records; therefore, the accuracy was only 57 to 64%. In the case of a balanced dataset, each of the classes has the same accuracy as trained on the same size of training data. Overall, SVM performed best on both of the datasets, followed by the stochastic gradient descent classifier; therefore, SVM can produce good results as text classification on top of contextual word embedding.

1. Introduction

The growth of scientific article publication has made finding important, relevant research difficult for researchers. Citations have long been studied for the identification of influential studies [1]. However, not all the citations within a research article play the same role. There may be different reasons for citing a research article, and therefore, the intensity of relatedness may vary. Moravcsik and Murugesan [2] argue that most of the references within articles are to understand the work and provide background knowledge about the research problem. Teufel et al. [3] have categorized the citations into three classes with a positive, weak, or neutral relationship with the citing paper. Jurgens et al. [1] have claimed that the citation maybe for six different reasons, and the strength of relevancy of these categories is different from each other.

Various attempts have been made in order to understand the reason and intent of a citation. The most recent techniques have used deep networks for reading the citation context of a citation [4–7]. They have set a window for extracting the citation context. The window boundaries typically contain the paragraph in which citation has been made. It may also include the sentences before and after it. An example of the citation context is given in Figure 1, being cited for comparison of the proposed methodology cited work.

Other approaches have utilized the bibliographic information of the research articles, which creates a network of citation nodes having edges of their mutual linking by citing [9]. These approaches reasonably find the relationship among the citation papers but usually fail to provide the reason for a citation as they give the same weight to each of



In Figure 6, the "Postive" score results of the proposed work are compared with Haider et al. [11] and Zafar et al. [39]. In this result, the top 5 best performer adverbs are selected from Haider et al. and Zafar et al.'s research studies.

Figure 1: A sample of citation context from a research article [8].

the references. Metainformation has been used extensively for citation intent extraction. The study based on text features is limited to the statistical similarity of the articles and normally does not study the internal context of those features [10]. New advances in natural language processing, especially word embedding, have made it possible to understand the text context and label them with a class of intent [11].

This paper has evaluated a number of classification methods after converting the text information to their numerical representation. We have used Association for Computational Linguistics-Anthology Reference Corpus (ACL-ARC) and Science Citation (SciCite) datasets, discussed in the next section, to extract text features related to citation records. The ultimate goal of classification was to find the citation intent based on our selected text features list. The experiments show that the linear support vector machine (Linear-SVM) classifier has performed well on both datasets. We also evaluated the classifiers for the prediction of individual citation intent class. The results show that the algorithms performed well, particularly for those class prediction where the training set was immense; for example, in the case of Linear-SVM, the "background" class has an F1 score of 86% while the other classes, including "future" and "extension," have 65% and 61%, respectively. The overall objectives of this study include the following:

(1) Understanding the impact of text features for citation intent classification while using contextual encoding
(2) Evaluating the results and comparing the classification models for citation intent labelling
(3) Understanding the impact of training set size classifiers' biasness towards the individual citation classes
(4) To exploit the authorship and titles for citation intent classification

The rest of the paper is organized as follows: in Section 2, we introduce existing citation intent classification methods and the number of labelled classes. Section 3 discusses the proposed study framework. The details of each of the steps are further discussed in the subsections of Section 3. Section 4 evaluates the classification models and compares the results. Finally, we conclude our study in Section 5.

2. Related Work

The citation intent, also called the reason for a citation or citation function, has long been studied to analyze the research article relationship. As each article has, on average, 40 references and with time, the number of referenced articles within a research paper is growing [12], it is essential to understand why a paper has been cited. This section discusses various attempts made to identify the citation reason.

Roman et al. [4] used contextual embedding for capturing the context of citation context. They used an automated method for annotating the unannotated dataset for citation intent and achieved good precision, recall, and F1 score. They also developed a vast dataset containing one million labelled citation context, named C2D-I. The author claimed the dataset as new a state-of-the-art dataset to design new citation intent approaches. C2D-I annotated the intent in three classes: background, method, and result. Although they could successfully develop a vast labelled dataset required for deep learning, they have not developed any recommender system to identify the citation reason. Their method was merely for the dataset annotation and not for the citation reason identification.

Hassan et al. [13] proposed a deep-learning-based approach for classifying the importance of a citation from a list of referenced papers. They argued that not all references have the same measure of relevancy. They used a Short-Term Long Memory- (LSTM-) based [14] deep-learning model to distinguish between important and unimportant citations. They also presented a classification model based on machine learning to select best-performing features using a Random Forest (RF) classifier [15]. The authors have listed 14 features of a citation context describing the reason for citation, apart from being an important or unimportant citation.

Cohan et al. [16] criticized predefined hand-engineered features such as linguistic patterns extracted from paper content and borrowed the idea of scaffolding from Swayamdipta et al. [17]. They assumed that better representations could be obtained directly from the data. They proposed a multitasking framework to incorporate knowledge from a paper structure. Their designed framework incorporates two tasks as structural scaffolding: (1) prediction of the section title and (2) predicting whether a citation is needed. Their scaffolding also predicts the citation intent of a citation as background, method, or result class. They also created a SciCite dataset out of 6,627 papers having 11,020 by crowdsourcing. The authors compared their model with the previous state-of-the-art Jurgens et al.'s [1] method for citation intent classification and achieved better results in terms of precision, recall, and F1 measures. The authors used pattern-based features including sequence of phases, parts of speech, lexical categories depicting the positive or negative sentiments, and specific categories such as words "we extended" and "compared with the previous state-of-the-art method." They borrowed the list of patterns from Simone Teufel [18] and extended it with newly identified patterns and categories. They further exposed topic-based features by arguing that a topic thematic framing can point out the citation function. For example, a citation context describing the methodology is more likely related to "uses" function, whereas a citation context providing some definition is from the "background" class.

They also explored the prototypical argument features and investigated a list of arguments that reflect a class of citations. For prototypical argument featuring, they identified frequently occurring arguments in syntactic positions. For example, the words "follow," "unfold," and "extend" frequently occur for "extend" class of citations. A vector representing the occurrence of an argument is created. The average of those occurrences decides the similarity of a citation towards a citation class. This study has used natural language processing features in detail to measure the citation reason and importance and has proved to be state-of-the-art research in this area. This study demonstrated that authors are sensitive to discourse structure and publication venue when citing a research paper.

Table 1 provides the list of citation Internet classes. The table also lists the dataset in which each of the classes is used. Some citation context examples are taken from these available datasets, which belong to those citation intent classes.

3. Proposed Study Framework

In this section, we discuss various steps of the proposed study, as depicted in Figure 2. The flow of the proposed study starts with the data processing and cleaning step, followed by converting text data to numeric representation. After converting the text data to numeric data, we apply different classification algorithms by feeding this data to the input layer to the classifiers. Finally, we gather the results and compare various evaluation measures for comparing the effects of classification algorithms. In the next step, we discuss the data preparation and preprocessing step in detail.

3.1. Data Preparation. The data preparation step starts with the extraction of data for our study. We used two state-of-the-art datasets ACL-ARC and SciCite. These datasets are publicly available and widely used for citation intent classification. ACL-Anthology Reference Corpus (ACL-ARC) is an Artificial Neural Network- (ANN-) based citation intent classification dataset [1,19]. The dataset has around 2,000 records. It has a number of features, including the citation context where in-text citation has been placed, citing and cited paper_id, which can be used to access the paper details using a web service, publication years, paper titles, author ids, extended context including more information on the in-text citation context, section number, section title, citation marker offsets, the sentence before the citation context, and finally, the most crucial feature of citation intent specifying the reason of a reference. The citation intent in the ARL-ARC dataset has six citation intent classes described in Table 2.

The second dataset that we have used is the SciCite dataset [3]. This dataset has achieved a 13 percent increase in the F1 score in comparison to the ACL-ARC. The dataset includes, along with some other unimportant features, the name of the section in which in-text citation is placed, citing and cited paper id, citation context, citation intent class, and the confidence level of the annotated citation intent class. The features included in the dataset are minimal, and only few match the features listed in ACL-ARC. The second state-of-the-art dataset contains the citation intent annotation in only three classes: background, method, and result. This dataset is five times larger than the ACL-ARC dataset, with over 9,159 instances with citation intent distribution listed in Table 2.

In order to keep the datasets persistent and for comparing and evaluating the results on both of these datasets, we made a balanced version of SciCite, which includes the missing required features for our study. From the name, it is clear that the balanced version of SciCite is a balanced one with an equal number of instances in each class. We used the Semantic Scholar API (https://api.semanticscholar.org/) by passing the citing and cited paper ID to extract the missing feature information.

3.2. Preparation of Textual Information. This study is based on the features selected from both of the datasets discussed in the previous section. Table 3 provides the list of features selected from both of the datasets for our study. The table also provides the reason for choosing these particular features as input for machine-learning classifiers.

The features contain information in text form and, therefore, need natural language processing preprocessing steps for making them ready to be taken as inputs. The following operations are performed as data preparation steps.

3.2.1. Tokenization. This task is used for breaking the paragraph or sentences into words by using whitespace or a special character as a token separator.

3.2.2. Stop Word Removal. Stop words include the words that frequently occur in text having no significant impact on the topic under discussion. They normally include parts of speech. Natural Language Toolkit (NLTK) [27] has defined a massive list of stop words in sixteen different languages.

3.2.3. Removing Punctuation and White Spaces. We extended the NLTK stop word list in Python by adding numbers and special characters to it while removing the stop words.

3.2.4. Case Conversion. Regardless of the position of the words in a sentence, we have changed the case of text to small so that the case of a text does not impact the meaning of a text.

3.2.5. Stemming. Kantrowitz et al. [28] have studied the effects of stemming on word embedding using TFIDF and have proved that it has remarkable results. It is a

Table 1: Description of citation intent classes with examples from respective datasets.

Intent class	Dataset	Description	Example
Background	(1) C2D-I (2) ACL-ARC (3) SciCite	Class of citations providing definitions, explanation of a topic or area	(1) The following four components have been identified as the key elements of a question related to patient care (Richardson et al., 1995) (2) The recent great advances in speech and language technologies have made it possible to build fully implemented spoken dialogue systems (Aust et al., 1995; Allen et al., 1996; Zue et al., 2000; and Walker et al., 2000)
Uses	ACL-ARC	The citing paper is using technique, dataset and results of a cited article	(1) We use the agreement checker code developed by Alkuhlani and Habash (2011) and evaluate our baseline (MaltParser using only CORE12), best-performing model (easy-first Parser using CORE12 + DET + LMM + PERSON + FN $*$ NGR g + p) and the gold reference (2) [. . .] We used the supervised WSD approach described in by Lee and Ng, 2002, for our experiments, using the naive Bayes algorithm as our classifier
Comparison	(1) ACL-ARC (2) SciCite In SciCite, this class is called Result	The citing author expresses the similarity of the method or results of the proposed method with a cited reference	(1) Similar to the work of Li et al., 2013, our summarization system consists of three key components: an initial sentence preselection module to select some important sentence candidates; the abovementioned compression model to generate n-best compressions for each sentence; and then, an ILP summarization method to select the best summary sentences from the multiple compressed sentences (2) Tateisi et al. also translated LTAG into HPSG (Tateisi et al., 1998) (3) We are going to make such a comparison with the theories proposed by J. Hobbs (1979, 1982) that represent a more computationally oriented approach to coherence and those of T.A. van Dijk and W. Kintch (1983), who are more interested in addressing psychological and cognitive aspects of discourse coherence
Motivation	ACL-ARC	The cited paper demonstrates the need for a new method, technique, or dataset	(1) This idea was inspired by Delisle et al. (1993), who used a list of arguments surrounding the main verb together with the verb's subcategorization information and previously processed examples to analyze semantic roles (case relations) xxxx. (2) Our motivation for generation of material for language education exists in work such as that of Sumita et al. (2005) and Mostow and Jang (2012), which deal with automatic generation of classic fill-in-the-blank questions
Extension	(1) ACL-ARC	The citing paper is extending the work or dataset of the referenced research	(1) We improve a two-dimensional multimodal version of LDA (Andrews et al., 2009) (2) Our work builds on earlier research on learning to identify dialogues in which the user experienced poor speech recognizer performance (Litman et al., 1999)
Method	(1) C2D-I (2) SciCite	Same as the extension class mentioned above	Same as the extension class mentioned above
Future	ACL-ARC	The cited research has potential use or extension in future work	(1) We perceive that these results can be extended to other language models that properly embed bilexical context-free grammars, such as, for instance, the more general history-based models used in the work of Ratnaparkhi, 1997, and Chelba and Jelinek, 1998. (2) Such a component would serve as the first stage of a clinical question answering system (Demner-Fushman and Lin, 2005) or summarization system (McKeown et al., 2003)
Important	Teufel [18]	The reference article is an important one and must be counted towards the main contribution or being extended by the citing article	(1) We use the nonprojective k-best MST algorithm to generate k-best lists (Hall, 2007), where $k = 8$ for the experiments in this paper (2) For better comparison with the work of others, we adopt the suggestion made by Green and Manning (2010) to evaluate the parsing quality on sentences up to 70 tokens long

Table 1: Continued.

Intent class	Dataset	Description	Example
UnImportant	Teufel [18]	The reference article may be merely for definition or discussion of how the area of research is important	(1) Typical letter-to-sound rule sets are those described by Ainsworth (1973), McIlroy (1973), Elovitz et al. (1976), Hurmicutt (1976), and Divay and Vitale (1997)

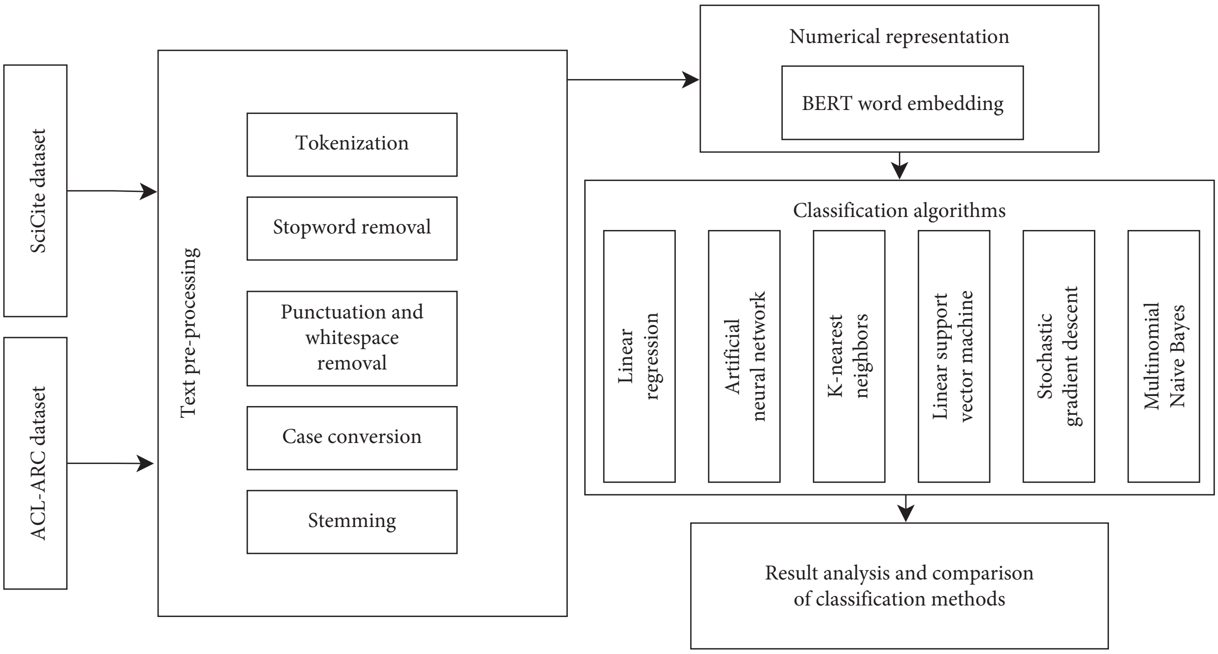


Figure 2: Framework of the proposed study for the citation content classification study.

Table 2: Distribution of records in citation intent classes in selected datasets.

SrNo.	Citation intent class	ACL-ARC dataset	SciCite dataset	Balanced SciCite dataset
1.	Background	1020	4,840	500
2.	Uses	365	—	—
3.	Comparison	344	1,109	500
4.	Motivation	98	—	—
5.	Extension	73	2,294	500
6.	Future Work	68	—	—

Table 3: The list of features selected for study and their availability in the selected datasets.

#	Feature name	Availability in ACL-ARC	Availability in SciCite	Remarks
1	Citing paper title	Available	Not available but can be extracted from Semantic Scholar API	The title has the most strong words which can express the purpose of a study and can be used to find the relevance of papers [20–22]
2	Cited paper title	Available	Not available but can be extracted from Semantic Scholar API	
3	Citing author	Available	Not available but can be extracted from Semantic Scholar API	The authors association can contribute to finding article relationship [23–26]
4	Cited author	Available	Not available but can be extracted from Semantic Scholar API	

language-specific task and converts words from the derived form to their root form. We have used the NLTK package for stemming the terms of our text data.

Once the text data are in a cleaner form, we need to convert the nlp_input to some numerical form as machine-learning algorithms required numerical representation of information for processing, discussed in the next section.

3.3. Numerical Representation of Text Data. The raw data in text format is converted into numerical representation such that similar words are closer to each other on the vector size. We used word embedding for numeric representation. Table 4 discusses various types of word embeddings along with their strengths and weaknesses. We have selected BERT word embedding as BERT is good in capturing the contextual information from a text and has been used by Roman et al. [4] for a similar task. BERT uses the transformers model [35, 36] for encoding the vector representation, using encoding-decoding architecture. We used Transformer libraries [37] for BERT implementation using Python language on the Kaggle platform (https://www.kaggle.com/).

3.4. Classification Models. Once the data has been converted to numeric representation, similar words are closed on the vector space. We are ready to feed this information to the citation classification model and evaluate the results to determine the best classification algorithm for citation intent class prediction. The classification methods assign predefined classes to the feature data. To define our problem, we consider our training dataset,

$$D = \{r_1, r_2, r_3, \ldots, r_n\}, \quad (1)$$

of records. Each record r_i is assigned a citation class c_i from

$$C = \{c_1, c_2, c_3, \ldots, c_n\}. \quad (2)$$

The task is to find the best classification method m, where

$$\begin{aligned} m&: D \longrightarrow C, \\ m(d) &= c, \end{aligned} \quad (3)$$

can assign an accurate citation intent to the new instance r. To study the accuracy of the classifiers, a number of classification algorithms have proved best for natural language processing tasks, listed in Table 5. The steps performed in this stage are listed below and depicted in Figure 3.

(1) The classification models were provided with the input parameters, listed in Table 5, from ACL-ARC and SciCite datasets. 80% of the records were provided as training data.

(2) We trained a model based on the input parameters, adjusting the input weights for the target class of citation intent.

(3) The trained model was then used for predicting 20% of the remaining records.

(4) The predicted citation class was checked with the actual class of the inputs.

(5) To guard against jumping to a conclusion without enough evidence, we calculated the average accuracy by repeating the experiments multiple times.

After setting the general guidelines and executing the steps discussed above, we performed an experiment and compared the selected machine-learning algorithms discussed in the next section.

4. Result Analysis and Comparisons

After training the models listed in Table 5, we performed experiments on the testing part of the datasets. In this section, we discuss the results of each model using precision, recall, and F1 measures. Precision counts positive predicted values and is the number of classes correctly identified. Recall is the fraction of actual classes identified. Increasing one decreases typically the other, and therefore, a harmonic mean of these two values is calculated given by the F1 measure. By evaluating the results against these measures, we want to see which model has performed well compared to the other models.

A multiclass confusion matrix is created using sklearn [44], NumPy, and seaborn libraries shown in Figures 4 and 5 for ACL-ARC and SciCite datasets. The confusion matrix clearly describes the number of true positive, false positive, false positive, and false negative predictions for each of the classes in the respective datasets. The calculation of precision is based on true positive and false negative parameters. The true positive is divided by the sum of true positive and false negative.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}. \quad (4)$$

A multiclass confusion matrix is given in Table 6 for the linear regression classifier on the ACL-ARC dataset. We used this table to present a sample calculation of precision, recall, and F1 score. The precision of a model is the average of the precisions of each of its classes. Thus, the precision of the linear regression classifier is calculated as follows:

$$\begin{aligned} \text{Precision}_{\text{background}} &= \frac{791}{791 + (42 + 34 + 14 + 15)}, \\ &= 0.88. \end{aligned} \quad (5)$$

Similarly,

$$\begin{aligned} \text{Precision}_{\text{uses}} &= 0.71, \\ \text{Precision}_{\text{comparison}} &= 0.69, \\ \text{Precision}_{\text{motivation}} &= 0.36, \\ \text{Precision}_{\text{extension}} &= 0.45, \\ \text{Precision}_{\text{future}} &= 0.61, \\ \text{Precision}_{\text{average}} &= \frac{0.88 + 0.71 + 0.69 + 0.36 + 0.45 + 0.61}{6} = 73\%. \end{aligned} \quad (6)$$

Thus, the average precision of the linear regression classifier is 73%. Similarly, the rest of the precisions are calculated for each classifier, given in Table 7 and 8 for

Table 4: Overview of word embedding techniques with their strengths and weaknesses.

#	Algorithm	Strengths	Weaknesses	Type of word embedding
1	TF-IDF [29]	(1) Vectors based on the occurrence of a word within a corpus and in the document are counted (2) Vector is proportional to the count of a word in a document and inverse to its count in other documents (3) Reducing the importance of common words frequently occurring, e.g., “while,” “but,” “the,” and “is” (4) Computing similarity is easy	(1) The similarity is merely based on the frequency of the words neglecting the semantic similarity (2) The size of a vector is large (3) Co-occurrence of words in a document is not recorded (4) Vectors are sparse (5) Synonyms are not considered (6) Polysemy words have a single vector. For example, apple is a fruit and Apple is a company; both have the same vector representation	Count based
2	Global Vectors (GloVe) [30], co-occurrence matrix [29]	(1) It is a hybrid method using a statistical matrix with machine learning (2) Records the appearance of a set of words in a corpus (3) Semantic similarity between King and Queen (4) Dimensionality reduction reduces the dimensions while producing more accurate vectors	(1) Costly in terms of memory, for recording co-occurrences of words	Count based
3	Word2Vec [31]	(1) Word analogies and word similarities are stimulated (2) Measures likelihoods of wordsxxxx (3) “King-man + woman = Queen,” which is a great feature of word embedding (4) Vectors can infer “king: man as queen: woman” (5) Input words mapped to target words (6) Probabilistic methods generally perform superior to deterministic methods [32] (7) Comparatively, small memory is consumed	(1) Training becomes difficult with the large size of the vocabulary (2) Polysemy words have an aggregated vector representation provided in CBOW, whereas in Skip-gram, they keep separate vectors	Prediction based
4	ELMO [33], Infersent [34], BERT [33]	Positioning embedding is incorporated, creating different vectors for the same word depending upon the position and context in a sentence/ paragraph	Contextualized embeddings require lots of computation	Prediction based

Table 5: Classification algorithms used for comparison for citation intent classification.

SrNo	Classifier
1.	Linear Regression (LR) [38]
2.	Artificial neural network (ANN) [39]
3.	K-nearest neighbors (KNN) [40]
4.	Linear support vector machine (linear-SVM) [41]
5.	Stochastic gradient descent (SGD) [42]
6.	Multinomial Naive Bayes (MNB) [43]

ACL-ARC and SciCite datasets, respectively. The second measure of evaluation is recall. Recall finds the proportion of actual positive correctly identified. To calculate recall, true positive is divided by the sum of true positive and false negative.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}. \tag{7}$$

The recall of linear regression on the ACL-ARC dataset is calculated as follows:

$$\begin{aligned}\text{Recall}_{\text{background}} &= \frac{791}{791 + (61 + 78 + 65 + 15 + 10)},\\ &= 0.78.\end{aligned} \tag{8}$$

Similarly,

$$\begin{aligned}\text{Recall}_{\text{uses}} &= 0.72,\\ \text{Recall}_{\text{comparison}} &= 0.77,\\ \text{Recall}_{\text{motivation}} &= 0.65,\\ \text{Recall}_{\text{extension}} &= 0.47,\\ \text{Recall}_{\text{future}} &= 0.57,\\ \text{Recall}_{\text{average}} &= \frac{0.78 + 0.72 + 0.77 + 0.65 + 0.47 + 0.57}{6} = 66\%.\end{aligned} \tag{9}$$

The average recall of linear regression using ACL-ARC is, thus, 66%.

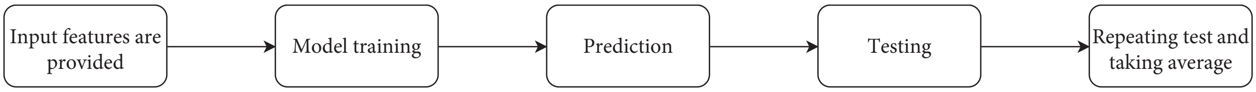
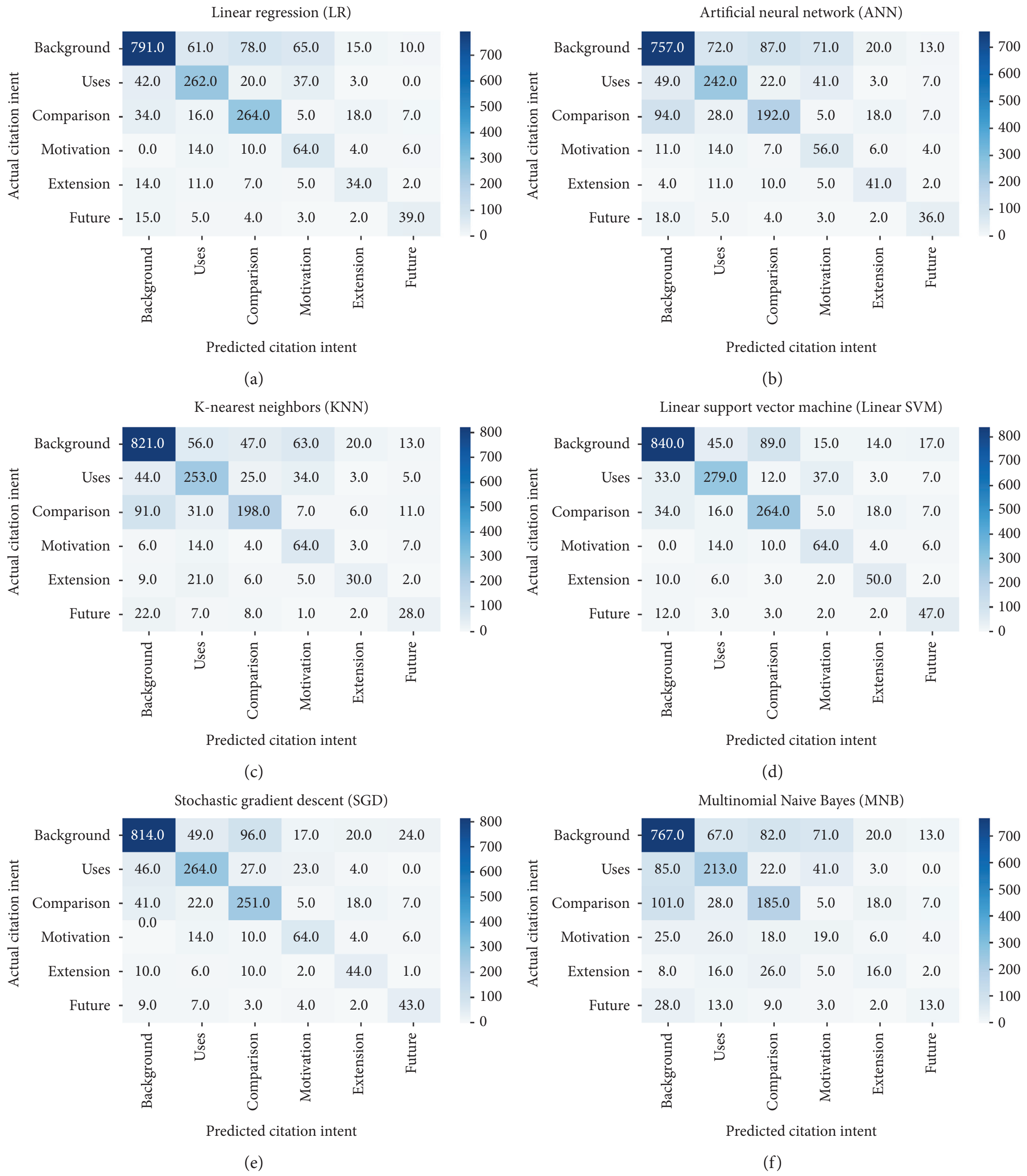

Input features are provided
Model training
Prediction
Testing
Repeating test and taking average

Figure 3: Steps for training the classification model.



Linear regression (LR)

	Background	Uses	Comparison	Motivation	Extension	Future
Background	791.0	61.0	78.0	65.0	15.0	10.0
Uses	42.0	262.0	20.0	37.0	3.0	0.0
Comparison	34.0	16.0	264.0	5.0	18.0	7.0
Motivation	0.0	14.0	10.0	64.0	4.0	6.0
Extension	14.0	11.0	7.0	5.0	34.0	2.0
Future	15.0	5.0	4.0	3.0	2.0	39.0

(a)

Artificial neural network (ANN)

	Background	Uses	Comparison	Motivation	Extension	Future
Background	757.0	72.0	87.0	71.0	20.0	13.0
Uses	49.0	242.0	22.0	41.0	3.0	7.0
Comparison	94.0	28.0	192.0	5.0	18.0	7.0
Motivation	11.0	14.0	7.0	56.0	6.0	4.0
Extension	4.0	11.0	10.0	5.0	41.0	2.0
Future	18.0	5.0	4.0	3.0	2.0	36.0

(b)

K-nearest neighbors (KNN)

	Background	Uses	Comparison	Motivation	Extension	Future
Background	821.0	56.0	47.0	63.0	20.0	13.0
Uses	44.0	253.0	25.0	34.0	3.0	5.0
Comparison	91.0	31.0	198.0	7.0	6.0	11.0
Motivation	6.0	14.0	4.0	64.0	3.0	7.0
Extension	9.0	21.0	6.0	5.0	30.0	2.0
Future	22.0	7.0	8.0	1.0	2.0	28.0

(c)

Linear support vector machine (Linear SVM)

	Background	Uses	Comparison	Motivation	Extension	Future
Background	840.0	45.0	89.0	15.0	14.0	17.0
Uses	33.0	279.0	12.0	37.0	3.0	7.0
Comparison	34.0	16.0	264.0	5.0	18.0	7.0
Motivation	0.0	14.0	10.0	64.0	4.0	6.0
Extension	10.0	6.0	3.0	2.0	50.0	2.0
Future	12.0	3.0	3.0	2.0	2.0	47.0

(d)

Stochastic gradient descent (SGD)

	Background	Uses	Comparison	Motivation	Extension	Future
Background	814.0	49.0	96.0	17.0	20.0	24.0
Uses	46.0	264.0	27.0	23.0	4.0	0.0
Comparison	41.0	22.0	251.0	5.0	18.0	7.0
Motivation	0.0	14.0	10.0	64.0	4.0	6.0
Extension	10.0	6.0	10.0	2.0	44.0	1.0
Future	9.0	7.0	3.0	4.0	2.0	43.0

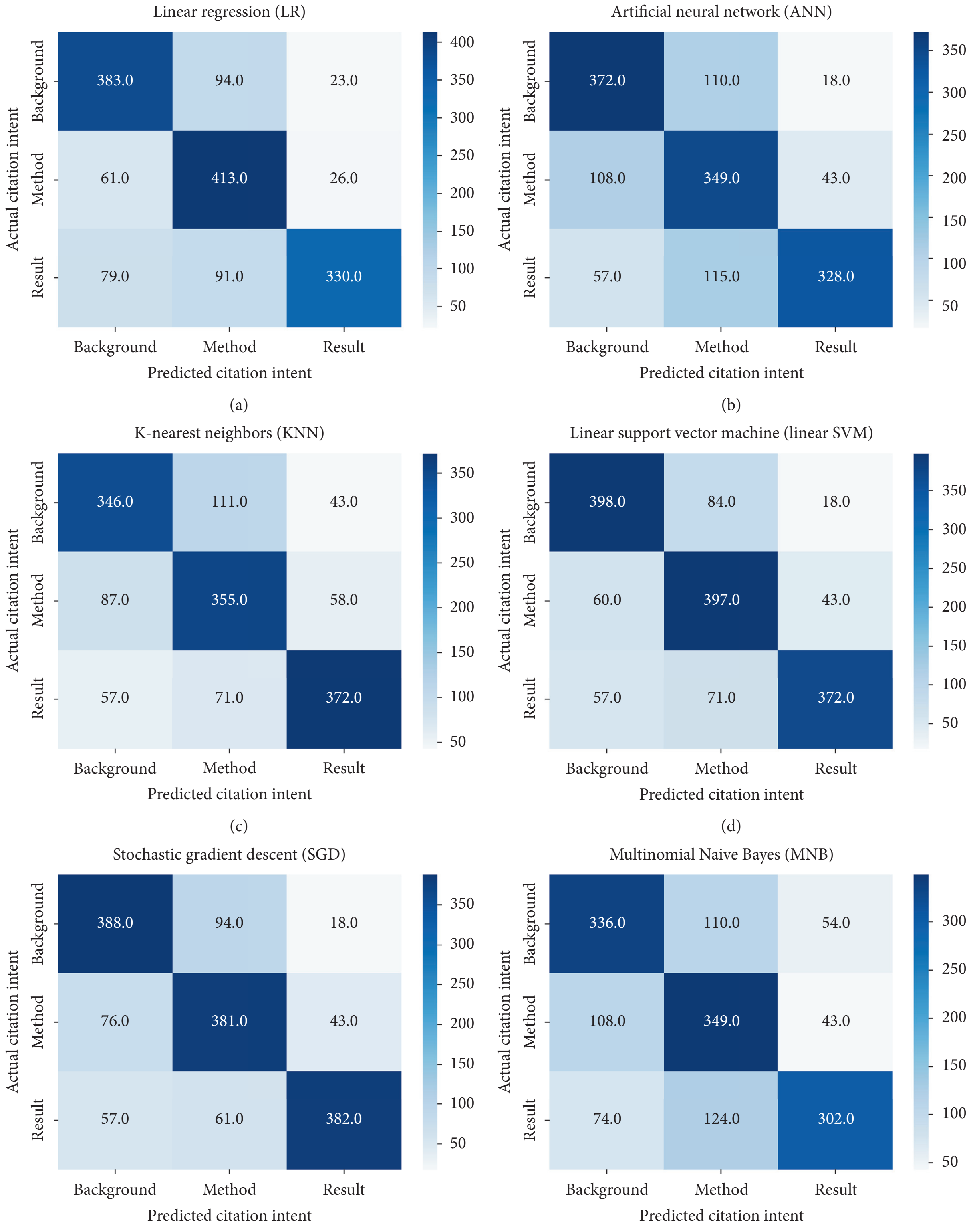
(e)

Multinomial Naive Bayes (MNB)

	Background	Uses	Comparison	Motivation	Extension	Future
Background	767.0	67.0	82.0	71.0	20.0	13.0
Uses	85.0	213.0	22.0	41.0	3.0	0.0
Comparison	101.0	28.0	185.0	5.0	18.0	7.0
Motivation	25.0	26.0	18.0	19.0	6.0	4.0
Extension	8.0	16.0	26.0	5.0	16.0	2.0
Future	28.0	13.0	9.0	3.0	2.0	13.0

(f)

Figure 4: Multiclass confusion matrix using the ACL-ARC dataset for each classifier.



(a) (b) (c) (d) (e) (f)

Figure 5: Multiclass confusion matrix using the balanced SciCite dataset for each classifier.

Table 6: Multiclass confusion matrix for linear regression using ACL-ARC.

Predicted label	Actual label					
	Background	Uses	Comparison	Motivation	Extension	Future
Background	791	42	34	0	14	15
Users	61	262	16	14	11	5
Comparison	78	20	246	10	7	4
Motivation	65	37	5	64	5	3
Extension	15	3	18	4	34	2
Future	10	0	7	6	2	39

Table 7: Comparison of precision, recall, and F1 score of various classifiers using the ACR-ARC dataset.

Classifiers	Background			Uses			Comparison			Motivation			Extension			Future		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
LR	0.88	0.78	0.83	0.71	0.72	0.71	0.69	0.77	0.73	0.36	0.65	0.46	0.45	0.47	0.46	0.61	0.57	0.59
ANN	0.81	0.74	0.78	0.65	0.66	0.66	0.61	0.56	0.58	0.40	0.57	0.40	0.46	0.56	0.50	0.52	0.53	0.53
KNN	0.83	0.80	0.82	0.66	0.70	0.68	0.69	0.58	0.36	0.37	0.65	0.47	0.47	0.41	0.44	0.42	0.41	0.42
L-SVM	0.90	0.82	0.86	0.77	0.77	0.77	0.69	0.77	0.73	0.51	0.65	0.57	0.55	0.68	0.61	0.59	0.68	0.64
SGD	0.88	0.80	0.84	0.73	0.73	0.73	0.63	0.73	0.68	0.56	0.65	0.60	0.48	0.60	0.53	0.53	0.63	0.58
MNB	0.76	0.75	0.75	0.59	0.59	0.59	0.54	0.54	0.54	0.13	0.19	0.16	0.25	0.22	0.23	0.33	0.19	0.24

Table 8: Comparison of precision, recall, and F1 score of various classifiers using the SciCite dataset.

Classifier	Background			Method			Result		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
LR	0.73	0.77	0.75	0.69	0.83	0.75	0.87	0.66	0.75
ANN	0.69	0.74	0.72	0.61	0.70	0.65	0.84	0.66	0.74
KNN	0.74	0.78	0.76	0.71	0.76	0.74	0.86	0.76	0.81
L-SVM	0.77	0.81	0.78	0.72	0.79	0.75	0.86	0.74	0.80
SGD	0.71	0.69	0.70	0.66	0.71	0.68	0.79	0.74	0.76
MNB	0.65	0.67	0.66	0.60	0.70	0.64	0.76	0.60	0.67

Precision and recall measures are always in tension, and increasing one results in decreasing the other. Therefore, a third measure called F1 score is used, which is a weighted average of the two previously calculated measures given by

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{10}$$

A sample calculation of the F1 for linear regression on the SciCite dataset is as follows:

$$\begin{aligned}
F1_{\text{background}} &= 2 \times \frac{\text{Precision}_{\text{background}} \times \text{Recall}_{\text{background}}}{\text{Precision}_{\text{background}} + \text{Recall}_{\text{background}}},\\
&= 0.83,\\
F1_{\text{uses}} &= 0.71\,,\\
F1_{\text{comparison}} &= 0.73,\\
F1_{\text{motivation}} &= 0.46,\\
F1_{\text{extension}} &= 0.46,\\
F1_{\text{future}} &= 0.59,\\
F1_{\text{Average}} &= \frac{0.83 + 0.73 + 0.46 + 0.46 + 0.59}{6},\\
&= 63\%.
\end{aligned} \tag{11}$$

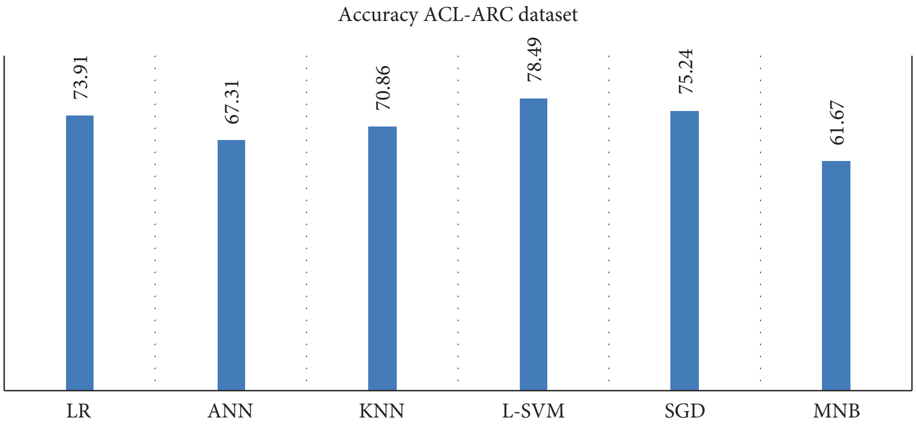


Figure 6: Accuracy of classifiers using the ACL-ARC dataset.

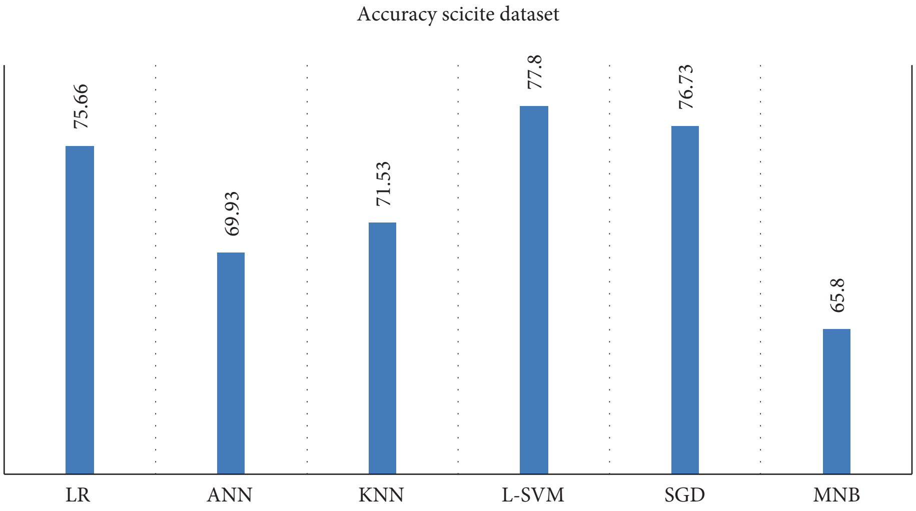


Figure 7: Accuracy of classifiers using the SciCite dataset.

The average F1 score of linear regression is only 63% using the ACL-ARC dataset. Although the F1 score of some of the citation intent classes is very high, in the case of background class, it is 83%, yet the overall F1 score of this classifier is significantly less. This is because of the unbalanced nature of the ACL-ARC dataset, as some of the other classes have minimal records in the dataset, and their training has not been performed very well.

Tables 7 and 8 provide a complete list of precision, recall, and F1 scores for each classifier. The overall accuracy of the classifiers is shown in Figures 6 and 7. Linear-SVM has the highest accuracy on both of the datasets, having 78.49% and 77.8%. Background class measures in the ACL-ARC dataset are much higher than the rest of the classes as the ACL-ARC is not a balanced dataset and, therefore, is biased towards the classes having a higher number of training records. Motivation, extension, and future classes have the least F1 score due to their small training data size, having less than 100 records in each of these cases. To further validate our conclusion, we observed that, in our balanced SciCite dataset, the F1 score is very closed for each of the classes, while the result class has the highest F1 score. The SGD classifier has the second-highest accuracy with little difference with the linear regression classifier.

5. Conclusions

Understanding the reason for a citation in a research article is crucial to investigate the essential related documents. Machine learning can perform well in classifying numeric metadata. Advances in natural language processing have made it possible to convert text data into a vector representation. The vectors can then be passed to classification algorithms to annotate the records in a scientific dataset. We have used BERT, a contextualized word representation, for converting text data to vectors. The classifiers were then evaluated, and two state-of-the-art datasets, ACL-ARC and SciCite, were used. The trained models performed well, especially in the case of our balanced version of SciCite. Linear SVM achieved an 86% F1 score on the “background” class where the training records were above 1000. In the case of citation intent classes where the number of training records were less than 100, SVM achieved only 57 to 64% F1 score. In the case of a balanced dataset, SVM and other algorithms did not have that much difference in the accuracy of the classifiers. This study has utilized only the text features from the dataset. In the future study, the meta- and NLP feature, consisting of text information, can both be combined to classify citation intent class.

Data Availability

The data are available upon request from the corresponding author.

Conflicts of Interest

The authors declare no conflicts of interest in this research study.

Acknowledgments

This work was supported in part by the Counterpart Service for the Construction of Xiangyang Science and Technology Innovation China Innovative Pilot City.

References

[1] D. Jurgens, S. Kumar, R. Hoover, D. McFarland, and D. Jurafsky, "Measuring the evolution of a scientific field through citation frames," *Transactions of the Association for Computational Linguistics*, vol. 6, 2018.

[2] M. J. Moravcsik and P. Murugesan, "Some results on the function and quality of citations," *Social Studies of Science*, vol. 5, no. 1, pp. 86–92, 1975.

[3] S. Teufel, A. Siddharthan, and D. Tidhar, "Automatic classification of citation function," in *Proceedings of the 2006 conference on empirical methods in natural language processing*, pp. 103–110, Sydney, Australia, July 2006.

[4] M. Roman, A. Shahid, S. Khan, A. Koubaa, and L. Yu, "Citation intent classification using word embedding," *IEEE Access*, vol. 9, pp. 9982–9995, 2021.

[5] A. Prasad, M. Kaur, and M.-Y. Kan, "Neural ParsCit: a deep learning-based reference string parser," *International Journal on Digital Libraries*, vol. 19, no. 4, pp. 323–337, 2018.

[6] D. R. Radev, P. Muthukrishnan, V. Qazvinian, and A. Abu-Jbara, "The ACL anthology network corpus," *Language Resources and Evaluation*, vol. 47, no. 4, pp. 919–944, 2013.

[7] M. E. Peters et al., "Deep contextualized word representations," 2018.

[8] U. A. Chauhan, M. T. Afzal, A. Shahid, M. Abdar, M. E. Basiri, and X. Zhou, *A Comprehensive Analysis of Adverb Types for Mining User Sentiments on Amazon Product Reviews*, Springer, Berlin, Germany, 2021, https://link.springer.com/article/10.1007/s11280-020-00785-z.

[9] R. Habib and M. T. Afzal, "Paper recommendation using citation proximity in bibliographic coupling," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 25, no. 4, pp. 2708–2718, 2017.

[10] H. Liu, X. Kong, X. Bai, W. Wang, T. M. Bekele, and F. Xia, "Context-based collaborative filtering for citation recommendation," *IEEE Access*, vol. 3, pp. 1695–1703, 2015.

[11] A. Dridi, M. M. Gaber, R. M. A. Azad, and J. Bhogal, "Scholarly data mining: a systematic review of its applications," *WIREs Data Mining and Knowledge Discovery*, vol. 11, no. 2, 2020.

[12] S. Milojević, "How are academic age, productivity and collaboration related to citing behavior of researchers?" *PLoS One*, vol. 7, no. 11, Article ID e49176, 2012.

[13] S.-U. Hassan, M. Imran, S. Iqbal, N. R. Aljohani, and R. Nawaz, "Deep context of citations using machine-learning models in scholarly full-text articles," *Scientometrics*, vol. 117, no. 3, pp. 1645–1662, 2018.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[16] A. Cohan, W. Ammar, M. van Zuylen, and F. Cady, "Structural scaffolds for citation intent classification in scientific publications," 2020, http://arxiv.org/abs/1904.01608.

[17] S. Swayamdipta, S. Thomson, K. Lee, L. Zettlemoyer, C. Dyer, and N. A. Smith, "Syntactic scaffolds for semantic structures," 2018, http://arxiv.org/abs/1808.10485.

[18] S. Teufel, "Argumentative zoning: information extraction from scientific text," Ph. D. thesis, University of Edinburgh, Edinburgh, Scotland, 1999.

[19] Z. Guo, K. Yu, Y. Li, G. Srivastava, and J. C.-W. Lin, "Deep learning-embedded social internet of things for ambiguity-aware social recommendations," *IEEE Transactions on Network Science and Engineering*, 2021.

[20] A. Shahid et al., "Insights into relevant knowledge extraction techniques: a comprehensive review," *The Journal of Supercomputing*, vol. 76, pp. 1–39, 2019.

[21] J. Zhang, K. Yu, Z. Wen, X. Qi, and A. Kumar Paul, "3D reconstruction for motion blurred images using deep learning-based intelligent systems," *Computers, Materials & Continua*, vol. 66, no. 2, pp. 2087–2104, 2021.

[22] L. Zhao et al., "Novel online sequential learning-based adaptive routing for edge software-defined vehicular networks," *IEEE Transactions on Wireless Communications*, 2020.

[23] A. Rexha, M. Kröll, H. Ziak, and R. Kern, "Authorship identification of documents with high content similarity," *Scientometrics*, vol. 115, no. 1, pp. 223–237, 2018.

[24] Y. Sun, R. Barber, M. Gupta, C. C. Aggarwal, and J. Han, "Co-author relationship prediction in heterogeneous bibliographic networks," in *in Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining*, pp. 121–128, Kaohsiung, Taiwan, July 2011.

[25] A. Y. Khan, A. S. Khattak, and M. T. Afzal, "Extending co-citation using sections of research articles," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 26, no. 6, pp. 3345–3355, 2018.

[26] K. Yu, L. Lin, M. Alazab, L. Tan, and B. Gu, "Deep learning-based traffic safety solution for a mixture of autonomous and manual vehicles in a 5G-enabled intelligent transportation system," *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[27] E. Loper and S. Bird, "NLTK: the natural language toolkit," 2002.

[28] M. Kantrowitz, B. Mohit, and V. Mittal, "Stemming and its effects on TFIDF ranking (poster session)," in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 357–359, Athens, Greece, July 2000.

[29] G.-H. Liu and J.-Y. Yang, "Image retrieval based on the texton co-occurrence matrix," *Pattern Recognition*, vol. 41, no. 12, pp. 3521–3527, 2008.

[30] J. Pennington, R. Socher, and C. D. Manning, "GloVe: global vectors for word representation," in *Proceedings of the*

Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543, October 2014, http://www.aclweb.org/anthology/D14-1162.

[31] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, http://arxiv.org/abs/1301.3781.

[32] M. Baroni, G. Dinu, and G. Kruszewski, "Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 238–247, Baltimore, MA, USA, June 2014.

[33] C. Jeong, S. Jang, H. Shin, E. Park, and S. Choi, "A context-aware citation recommendation model with BERT and graph convolutional networks," 2020, http://arxiv.org/abs/1903.06464.

[34] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 670–680, September 2017, https://www.aclweb.org/anthology/D17-1070.

[35] A. Vaswani et al., "Attention is all you need," 2017, http://arxiv.org/abs/1706.03762.

[36] Z. Guo, K. Yu, A. Jolfaei, A. K. Bashir, A. O. Almagrabi, and N. Kumar, "A fuzzy detection system for rumors through explainable adaptive learning," *IEEE Transactions on Fuzzy Systems*, 2021.

[37] T. Wolf et al., "Transformers: state-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, October 2020, https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[38] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, John Wiley & Sons, Hoboken, NJ, USA, 2021.

[39] A. Alghoul, S. Al Ajrami, G. Al Jarousha, G. Harb, and S. S. Abu-Naser, "Email classification using artificial neural network," *International Journal of Academic Engineering Research (IJAER)*, vol. 2, no. 11, pp. 8–14, 2018.

[40] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[41] S. Dumais, "Using SVMs for text categorization," *IEEE Intelligent Systems*, vol. 13, no. 4, pp. 21–23, 1998.

[42] S. Ruder, "An overview of gradient descent optimization algorithms," 2016, http://arxiv.org/abs/1609.04747.

[43] S. Xu, Y. Li, and Z. Wang, "Bayesian multinomial naïve bayes classifier to text classification," in *Advanced Multimedia and Ubiquitous Engineering*pp. 347–352, Singapore, 2017.

[44] F. Pedregosa et al., "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.