applied sciences

MDPI

Article

Citation Context Analysis Using Combined Feature Embedding and Deep Convolutional Neural Network Model

Musarat Karim [1], Malik Muhammad Saad Missen [1,*], Muhammad Umer [1], Saima Sadiq [2], Abdullah Mohamed [3] and Imran Ashraf [4,*]

[1] Department of Computer Science and Information Technology, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan; musarat.km11@gmail.com (M.K.); umersabir1996@gmail.com (M.U.)
[2] Department of Computer Science, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan 64200, Pakistan; s.kamrran@gmail.com
[3] Research Centre, Future University in Egypt, New Cairo 11745, Egypt; Mohamed.a@fue.edu.eg
[4] Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, Korea
* Correspondence: saad.missen@gmail.com (M.M.S.M.); imranashraf@ynu.ac.kr (I.A.)

Abstract: Citation creates a link between citing and the cited author, and the frequency of citation has been regarded as the basic element to measure the impact of research and knowledge-based achievements. Citation frequency has been widely used to calculate the impact factor, H index, i10 index, etc., of authors and journals. However, for a fair evaluation, the qualitative aspect should be considered along with the quantitative measures. The sentiments expressed in citation play an important role in evaluating the quality of the research because the citation may be used to indicate appreciation, criticism, or a basis for carrying on research. In-text citation analysis is a challenging task, despite the use of machine learning models and automatic sentiment annotation. Additionally, the use of deep learning models and word embedding is not studied very well. This study performs several experiments with machine learning and deep learning models using fastText, fastText subword, global vectors, and their blending for word representation to perform in-text sentiment analysis. A dimensionality reduction technique called principal component analysis (PCA) is utilized to reduce the feature vectors before passing them to the classifier. Additionally, a customized convolutional neural network (CNN) is presented to obtain higher classification accuracy. Results suggest that the deep learning CNN coupled with fastText word embedding produces the best results in terms of accuracy, precision, recall, and F1 measure.

Keywords: in-text citation; citation context analysis; deep learning; convolutional neural network; word embedding

check for updates

Citation: Karim, M.; Missen, M.M.S.; Umer, M.; Sadiq, S.; Mohamed, A.; Ashraf, I. Citation Context Analysis Using Combined Feature Embedding and Deep Convolutional Neural Network Model. *Appl. Sci.* **2022**, *12*, 3203. https://doi.org/10.3390/app12063203

Academic Editors: Jae-Hoon Kim, Kichun Lee and Elisa Quintarelli

Received: 12 February 2022
Accepted: 17 March 2022
Published: 21 March 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In pedagogical research, research articles are not presented as standalone work but embedded in the related literature. Citation is used to create a link for research articles with other scientific research works. Citations referred to as expressions are used to acknowledge other scientific works and the references are referred to as identifiers and are used to represent the cited work [1]. Citations not only express the sentiments of the authors citing the paper but also elaborate the importance of the cited work [2]. Citation identifies that a document or a piece of information is taken from another published research work. Citation plays a significant role in evaluating the impact of the research such as peer judgment and impact factor calculation [3]. In recent times, all types of citations are being considered equally for calculating different research performance indicators such as h-index, i10-index, g-index, etc. However, studies state that all citations are not the same [4], and quantitative measures alone cannot be used for a fair evaluation of research impact [5]. The qualitative aspect of citations should also be considered to calculate the impact of citation.

The quantitative measure is the frequency count of citations, and it is not a suitable measure [6], whereas the qualitative measure involves the measure of the quality and

polarity of a cited text, and is more critical to measure [7]. Therefore, different researchers have considered different measures to evaluate the quality of research, such as citation reason [8], citation sentiment [9], citation frequency, and text similarity measures [10]. Sentiment analysis is a technique for identifying the opinions, sentiments, thoughts, and views indicated in the text. Identifying the sentiments of the citation is very useful and interesting, and it helps researchers determine the quality of a scientific work [11] by analyzing the sentiment of citations for a particular work. With the increase in the amount of research, the need to analyze the sentiments of the citations in scientific papers also increases. In recent past years, the sentiment analysis of scientific citations is gaining increased attention from the research community.

The main aim of the citation sentiment analysis is to identify the sentiment polarity that the author carries towards the cited paper [12]. Analysis of the citation sentiment will also be helpful in new applications for automatically evaluating the impacts of journals and individuals by the use of the citations. It will also open up many applications in bibliometric and bibliographic searches [13]. Citation sentiment analysis is important because it helps with identifying the impact of the research, the importance of the research, the areas where no significant research has been made, and the limitations of a certain approach. It is also helpful for ranking the research papers, determining the most used methodology or a new approach proposed by the researcher, the reason for the research, the impact of the research, the opinions of the authors towards other authors' research, and much more [9]. For citation sentiment analysis, various machine learning approaches, along with different feature extraction techniques, have been utilized. Statistical techniques that increase the discriminative capability of classifiers, such as principal component analysis (PCA), have been applied for feature selection [14].

The most important application of citation sentiment analysis is for bibliometric measures. The citation sentiment analysis helps enhance the bibliometric measures. The previous way to study an article's impact is by counting the number of times that an article has been cited. However, citation sentiment analysis can be used to give weights to each citation text, considering the sentiments of the citations [15]. In most cases, the sentiments expressed in the citation text are hidden, and it is difficult to identify the polarity (positive, negative, or neutral) of the sentiments [16]. When it comes to humans, it is easy for them to read the citation text and identify the sentiment expressed in the citation text, but when it comes to training a model to automatically predict sentiments' polarity, it becomes a difficult and challenging task [17] because there exist many techniques to train the model using a dataset and to perform sentiment prediction. For the most part, the sentiment polarity of the cited text seems to be neutral, with negative or positive sentiments hidden [18]. For hidden sentiment analysis, several approaches can be used, such as lexical analysis, feature-based extraction, structure-based sentiment analysis, etc. This study leverages the machine learning approach for the automatic sentiment classification of in-text sentiments and makes the following key contributions:

- A framework is presented for in-text citation sentiment analysis based on a customized convolutional neural network (CNN) coupled with the combined word embedding approach. In addition, fastText, fastText subword, and global vectors for word representation (GloVe) are also analyzed individually for their efficacy for sentiment analysis;
- The unbiased sentiment of various citations is predicted to highlight the importance of a research work. By classifying the sentiment of in-text citations into positive, negative, or objective sentiments, the research significance of a work can be evaluated;
- An evaluation of the proposed framework is carried out by comparing its result with state-of-the-art models used in text classification, such as random forest (RF), stochastic gradient descent (SGD), logistic regression (LR), a voting ensemble that combines LR and SGD (VC(LR+SGD)), and long short-term memory (LSTM);
- Performance analysis is carried out with recent state-of-the-art approaches to show the significance of the proposed framework. For validating the results, the "clinical

trail" citations dataset is used. In addition, we also collect our own dataset to analyze the quality of selected science articles by using citation sentiments.

The rest of the paper is organized as follows. Section 2 discusses important works related to the current study. Section 3 presents an overview of the methodology adopted for the current research as well as a detailed description of the dataset and models used for experiments. Results are discussed in Section 5, while the conclusion and future work are provided in Section 7.

2. Related Work

The analysis of the citation sentiments for scientific papers is a newly emerging problem. In order to automatically identify the sentiment polarity (positive, negative, or neutral) in the citation, an experiment was conducted by Awais Athar using the existing features, which include; n-grams, 3-grams, lexical features, dependency relations, negations, and sentence splitting features on the citation sentiment corpus. The results shown in the research were that dependency relations and the 3-grams show the best results of all [9]. The objective of the citation sentiment analysis is to determine different patterns; with the bulk data availability, it is a need to extract useful information. The main aim was to automate the method for obtaining the sentiment polarity—either positive, negative, or neutral. Sentiment analysis is the need to determine the citation's sentiment polarity. Different trends were discussed, and an effective approach was recommended [18].

Citation sentiment analysis is performed to analyze the sentiments of the author towards the cited paper. A survey to analyze the sentiments of scientific citations was made by the authors, in which an introduction to the sentiment analysis process was given, the challenges faced by the previously proposed methods were discussed and presented, and an analysis was made. The main focus of the research was to identify the challenges that the recently proposed methods are facing, and how effective they are in citation sentiment analysis; the used classifications were also analyzed. It was concluded that machine learning is the most common method used for the analysis of the citation sentiments in scientific papers, and it was also seen that, because of the limitation of this method, deep learning methods can be used to effectively analyze the sentiment polarity of citations in scientific papers [15].

Citation sentences were used to analyze citation sentiments. The existing citation corpus was used. The corpus comprises 8736 citation sentences. To clean the data, different normalization rules were used. The classification was performed on six different classifiers of machine learning. Different evaluation parameters, such as accuracy, precision, recall, and F1 score, were used to evaluate the accuracy of the models. To improve the model's accuracy, different approaches were recommended, such as n-gamming, stop words removal, lemmatization, stemming, etc. The accuracy was improved by 9% using these methods [19].

With the increase in the amount of online data, especially textual data, sentiment analysis has become an emerging field. The classification of sentences or documents is completed by either considering the objective or the subjective textual data. Then, the classifications are collected and the sentiment analysis technique is applied to it to check the sentiment polarity (positive, negative, or neutral) of the text to predict the sentiments of the citation text. Several text preprocessing techniques are available that are used to preprocess the textual data to obtain a dataset that is more clear and concise and that comprises only useful features, excluding the useless features. The polarity of the cited text is analyzed by using the machine learning classifiers [20,21]. However, many state-of-the-art approaches are present, and are used to evaluate the accuracy of the models and to extract the sentiments expressed in the cited text. A method for determining the in-text citations using the citation sentiment analysis methodology was proposed, in which different machine learning models were used to first train the dataset. Then, testing was performed on unseen datasets to check the accuracy of the machine learning models in the case of in-text citations. Random forest, support vector machine, and kernel logistic regression models were used in the research to evaluate the accuracy [22]. A literature review reveals that many machine

learning models and manual feature engineering techniques have been analyzed by many researchers. However, deep learning models and word embedding techniques have not yet been analyzed for sentiment analysis of in-text citations.

3. Proposed Research Methodology

3.1. Dataset Description

This study uses two datasets for experiments, including the "citation sentiment corpus" dataset [9], acquired from the ACL corpus, and the "clinical trials" data from [16]. Dataset-1 includes 8736 in-text citations that are annotated manually. The dataset has four attributes: "Source_Paper ID", which is the ID of the paper from which text has been taken (the citing paper), "Target_Paper ID", which is the ID of the target or the cited paper, "Sentiment" with "p", "n", and "o" values, where "p" indicates positive, "n" indicates negative, and "o" indicates objective (neutral) sentiment, and "Citation_Text", which represents the citation text. A detailed description of the dataset is presented in Table 1. Similar to dataset-1, dataset-2 contains in-text citations for 285 articles from clinical trial articles. It has a total of 4182 citations that are annotated for citation sentiment, regarding positive, negative, and neutral sentiments.

Table 1. Details of citation sentiment corpus.

Attribute	No. of Records
Dataset-1	
No. of citation texts	8736
No. of papers	194
No. of positive instances	829
No. of negative instances	280
No. of objective instances	7627
Dataset-2	
No. of citation texts	4182
No. of papers	285
No. of positive instances	702
No. of negative instances	308
No. of objective instances	3172

3.2. Data Preprocessing

Preprocessing refers to transforming the redundant, missing, unnecessary, and inconsistent data into an appropriate format that is suitable for the models' training. Various steps are performed during the preprocessing phase to improve the suitability of the raw data and elevate the performance and efficiency of the models. Preprocessing steps of stopword removal, conversion to lowercase, and tokenization have been performed with the help of a natural language tool kit (NLTK) and Keras libraries in Python.

3.3. Word Embedding Techniques

For training machine learning and deep learning models, textual data is represented as vectors. It is an essential step in the natural language processing task. During recent years, word embedding feature engineering techniques have gained increased popularity regarding their use with prediction approaches. Three word embedding techniques have been used in this study for citation sentiment analysis, such as fastText, fastText subword, and GloVe.

3.3.1. fastText

Vector representation by word embedding has been used in various natural language processing (NLP) tasks. Generally, pre-trained word embedding that is trained on texts such as Wikipedia and Google News predicts the word context in an unsupervised way. They consider the words to appear close to each other and have similar contexts. The fastText [23] technique is created by Facebook's FAIR Research Lab. It contains 2 million word vectors (600 billion tokens) having 300 dimensions. fastText embedding is a good choice for vector representation because of the word difficulty detection feature, which uses morphological information. This capability makes it possible to improve the results of text classification by generalizing them well. Vectors of fastText word embedding are obtained by the sum of n-grams that make it possible to generate a vector for unfamiliar words.

3.3.2. fastText Subword

fastText subword [24] benefits the training process by sharing common roots in words. It contains 2 million word vectors (600 billion tokens) having 300 dimensions, and uses subword information for training. fastText subword consists of millions of vectors that are trained on common roots or tokens. It provides support by breaking words into subwords or by joining subwords into a single word. For example, individually represented, "for" and "give" can be represented as "forgive" in order to learn dictionary-level representation. Character level embedding helps with representing misspelled and slang words.

3.3.3. GloVe

GloVe is a word embedding technique based on an unsupervised learning model for obtaining vectors [25]. GloVe has 300 dimensions and contains 840 billion crawl words. As the name suggests, it captures the global and local features of the corpus. It is non-contextual and maps words according to the semantic similarity between them, presenting the word vector space of linear substructures. It is trained on the global aggregation of word co-occurrence. GloVe is a log bilinear model and works on the probability of words' co-occurrence and by extracting meaning from it. It generates a word context matrix by the factorization technique.

3.3.4. Combining Features

This study employs a combination of three word embedding techniques, namely, fastText, fastText subword, and GloVe. These are individually trained word embedding techniques that help the deep learning models obtain more accurate predictions. First, a deep learning model is trained on three word embedding techniques individually, and then by blending these techniques, with a ratio of 33% for fastText and Glove, and of 34% for fastText subword.

3.4. Dimensionality Reduction

Combining multiple word embedding techniques increases the number of features and duplications of features that increase the excessive burden on the training of the classifier. Dimensionality reduction, using an appropriate feature selection technique, can solve this problem. PCA uses a linear transformation to reduce the number of features, and has been extensively used for classification. The dataset reduced by PCA contains characteristics such as original data, and the principal component is computed with the help of the covariance matrix. For the current study, PCA is used to select the top 2000 features out of a total of 3400 features for experiments.

3.5. Modeling Methods

CNN is a deep neural network that maneuvers the computational complexity of large-sized data [26]. CNN is an efficient neural network model and learns complex features with the help of convolution, nonlinear activation, and dropout and pooling layers [27]. It was designed for image-related tasks, such as image segmentation and image classification [28].

In CNN, training is performed in an end-to-end fashion, which makes it more efficient. To encode semantic information, fully connected layers are utilized at the end of the model. It is a feed-forward network where filters are applied to the output of the previous layer to map features. The main components of the CNN model are convolutional layers, pooling or sub-sampling layers, a flatten layer, an activation function, dropout, and a fully connected layer. Features are extracted by convolutional layers, and then the output of the convolutional layers is fed to the fully connected layers. The pooling layer reduces the features mapped by convolutional layers to reduce overfitting probability. Pooling can be a max or average layer, where the max-pooling layer chooses sharp features as compared to the average pooling layer. The flatten layer converts the data into an array so that it can be fed to the fully connected layer. This study utilizes the rectified linear unit (ReLU) as an activation function:

$$y = max(0, i), \tag{1}$$

where y represents the activation output and i represents the input. Convolution layers extract local and high-level features by assigning weights to the kernel during the training phase. CNN has been widely used in disease diagnosis. In binary classification, the cross-entropy error is used as a loss function; this has also been used in this study. It is computed as:

$$cross - entropy = -(i * log(p) + (1 - i)log(2 - p)), \tag{2}$$

where i represents the indicator of class labels, a log is a natural log, and p represents the probability that is predicted.

As CNN is a modification of the backpropagation algorithm; therefore, sigmoid is utilized as the error function for output. The CNN model generates output as three neurons for each case of the target class. CNN has been regarded as a robust model for classification tasks in the medical field. CNN has been utilized by many researchers for various classification tasks, such as lung-disease classification [29], the segmentation of brain tumors [30], and X-rays of the chest [31]. In previous literature, CNN has also been analyzed for text categorization, such as text sentiment analysis [32], text summarization [33], and text report classification [34]. CNN has been employed to detect vision-threatening eye diseases using medical reports in [35].

Figure 1 presents the proposed framework for the sentiment analysis of in-text citations. The CNN model used in this study has been optimized through a customized structure in terms of the number of layers, number of neurons, and optimizer choice, etc. Details for the CNN architecture are provided in Table 2. In addition to CNN, several well-known machine learning models have been used for the task at hand, including RF, SGD, LR, and LSTM for comparison. The performance of these models is optimized by fine-tuning different hyperparameters, a list of which is given in Table 2.

Table 2. The layer structure and hyperparameters of the learning models.

Model	Structure
CNN	Conv (7 × 7, @64), Max pooling (2 × 2), Conv (7 × 7, @64), GlobalMax pooling (2 × 2), Dropout (0.5), Dense (32 neurons), Softmax (3), Categorical cross entropy
LSTM	LSTM (100 neurons), Dropout (0.5), Dense (32 neurons), Softmax (3)
RF	n_estimator = 200, max_depth = 30, random_state = 52
SGD	penality = “I2”, loss = “log”
LR	penalty = “I2”, solver = “, lbfgs”
VC	Voting = “soft”

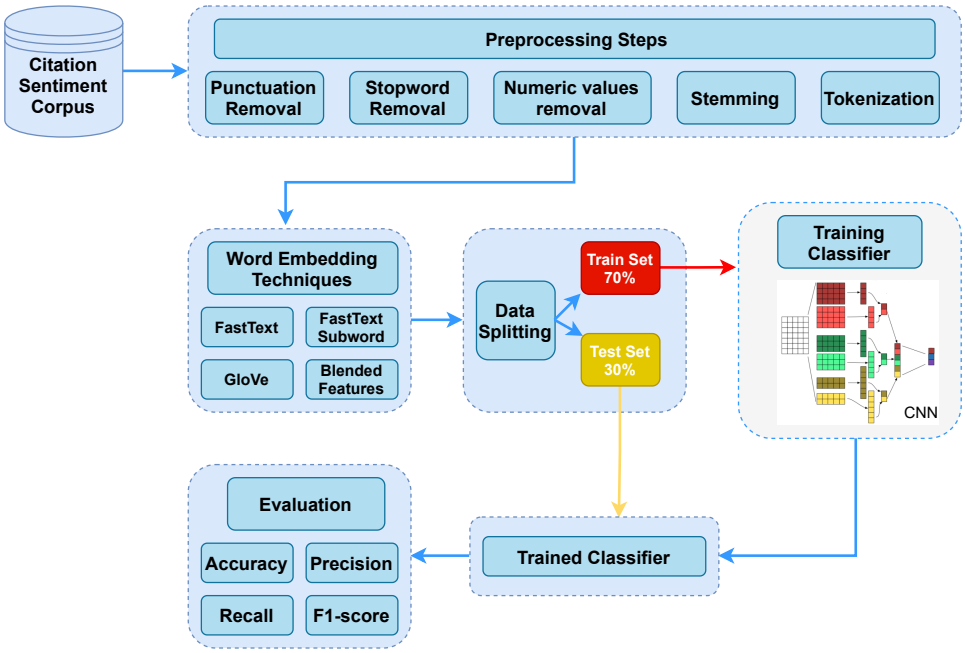

Figure 1. The architecture of the proposed methodology for in-text citation sentiment analysis.

4. Experiment Setup

This section describes the experimental setup and the machine learning models used for the experiments.

4.1. Models Used for Experiments

This study utilizes several machine learning and deep learning models on the selected dataset to compare their performance with the proposed approach. A brief description of these models is provided for completeness.

4.1.1. Random Forest

RF is a tree-based machine learning model that integrates the aggregated results acquired by fitting many decision trees on randomly selected training data samples [36]. Each decision tree in RF is generated on indicators such as the Gini index and information gain to select a root node. It is a meta-estimator that can be utilized for regression as well as classification tasks [7]. RF shows good results for text classification tasks [37,38].

4.1.2. Logistic Regression

LR is a statistical model that processes the mapping between a given set of input features by sigmoid function with a discrete set of target variables by approximating the probability. The sigmoid function is an S-shaped curve that restricts the probabilistic value between target variables [39]. It works efficiently for text classification tasks.

4.1.3. Stochastic Gradient Descent

SGDC joins various binary classifiers and has been extensively applied on a huge dataset. Its working mechanism is quite similar to the regression method and is easy to implement and interpret [40]. Hyperparameter values for SGD need to be correct to obtain correct results. SGD is sensitive in terms of feature scaling.

4.1.4. Long Short-Term Memory Network

A recurrent neural network (RNN) is a feed-forward deep neural network model that faces vanishing gradient problems and loss of information when dealing with long sequences of information. LSTM is an extended form of RNN. LSTM saves information

and deals with long sequences effectively with the help of memory cells and three gates. It uses structured gates to add or forget information to control memory cells. A forget gate is used to decide which information to remove [13]. The sigmoid function is used for this purpose; if the output is 1, information is remembered, and if the output is 0, it forgets. It is performed based on the current state and previous state.

4.2. Evaluation Parameters

Machine learning and deep learning models are evaluated with the help of different parameters. Accuracy, precision, recall, and F1 score have been used to show the performance comparison of different models in this work.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (3)$$

$$Precision = \frac{TP}{TP + FP}, \quad (4)$$

$$Recall = \frac{TP}{TP + FN}, \quad (5)$$

$$\text{F1 } Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \quad (6)$$

where TP, TN, FP, and FN represent the true positive, true negative, false positive, and false negative, respectively, and are extracted from the confusion matrix of each classifier.

5. Results

All the experiments are performed using a 2 GB Dell PowerEdge T430 graphical processing unit on a 2× Intel Xeon 8 Core 2.4 Ghz machine with 32 GB DDR4 random access memory (RAM). A Jupyter notebook environment is used to perform experiments in Python programming language with Anaconda. Machine learning models and deep learning models are implemented using sklearn, Keras, and Tensorflow.

5.1. Comparison of Predictive Performance of Models Using Dataset-1

Extensive experiments have been carried out for textual sentiment analysis. Efforts are underway to develop an efficient method for in-text citations' sentiment analysis. Machine learning models and deep learning models used in the experiments are CNN, LSTM, RF, SGD, LR, and a voting classifier that combines LR and SGD. Word embedding techniques, namely, fastText, fastText subword, GloVe, and their combination are investigated for citation sentiment analysis.

5.1.1. Experiments Using fastText

At first, the models are trained on fastText word embedding, and the results are presented in Table 3. All machine learning models show almost similar results and achieve an 81% F1 score using fastText for the sentiment analysis of in-text citations. SGD achieves 86.89% accuracy, which is the highest among all machine learning models. RF achieves the highest precision score, 79%, which is 3% greater than the precision achieved by other models. SGD, LR, and VC (LR + SGD) achieve 87% recall. For deep learning models, LSTM achieved the lowest result, with 86% accuracy, 70% precision, 74% recall, and 72% F1 score. However, the customized CNN achieves the best result, with the highest accuracy (89%), precision (87%), recall (85%), and F1 score (86%) for citation sentiment analysis.

Table 3. Classification results of classifiers using fastText.

Model	Accuracy	Precision	Recall	F1 Score
CNN	89%	87%	85%	86%
LSTM	86%	70%	74%	72%
RF	86.21%	79%	86%	81%
SGD	86.89%	76%	87%	81%
LR	86.78%	76%	87%	81%
VC (LR + SGD)	86.61%	76%	87%	81%

5.1.2. Experiments Using fastText subword

A separate set of experiments is performed using the machine learning and deep learning models with the fastText subword word embedding. Table 4 presents the performance comparison of models using fastText subword for the sentiment analysis of in-text citations. A results comparison reveals that the CNN, LSTM, RF, SGD, and LR models show a marginal improvement in their performance, except for the voting classifier. The improvement is observed in terms of better accuracy, precision, recall, and F1 score. CNN achieves the highest precision value, of 85%, while LSTM achieves 84% precision. The voting classifier achieves the lowest accuracy of 85.52% and the highest recall of 86%. CNN and RF achieve the highest F1 score of 83% each.

Table 4. Classification results of classifiers using fastText subword.

Model	Accuracy	Precision	Recall	F1 Score
CNN	87%	85%	82%	83%
LSTM	87%	84%	80%	81%
RF	87.35%	82%	87%	83%
SGD	87.41%	76%	87%	82%
LR	87.24%	76%	87%	81%
VC (LR + SGD)	85.52%	76%	86%	81%

5.1.3. Experiments Using GloVe Features

Next, models are trained on GloVe word embedding for citation sentiment analysis. The classification results of the classifiers are shown in Table 5. It can be observed that CNN surpassed other models with 86% accuracy, 88% precision, 89% recall, and 88% F1 score for citation sentiment analysis, while LSTM achieved 84.4% accuracy, which is even lower than the machine learning models. The voting classifier has shown the lowest results, with 81.19% accuracy and 77% F1 score. Machine learning models have shown almost similar results using GloVe for in-text citations' sentiment analysis.

Table 5. Experimental results using the GloVe features.

Model	Accuracy	Precision	Recall	F1 Score
CNN	86%	88%	89%	88%
LSTM	84.40%	78%	74%	76%
RF	85.84%	81%	86%	80%
SGD	85.53%	73%	86%	79%
LR	85.31%	76%	85%	79%
VC (LR + SGD)	81.19%	75%	81%	77%

5.1.4. Experiments Using Combined Features

Finally, models are trained by combining all three feature embedding techniques, fastText, fastText subword, and Glove, with a ratio of 33%, 34%, and 33%, respectively. Then, PCA is applied to extract important features by reducing dimensions. It can be observed that combining multiple word embedding techniques significantly improves the models' performance, as presented in Table 6. The proposed CNN, when used with combined features, obtains the highest results with 93.47% accuracy, 94.24% precision, 96.18% recall, and 95% F1 score for the sentiment analysis of in-text citations.

Table 6. Experimental results using the combined features.

Model	Accuracy	Precision	Recall	F1 Score
CNN	93.47%	94.24%	96.18%	95%
LSTM	89.24%	91.36%	93.21%	92.28%
RF	87.87%	89%	92%	90.05%
SGD	88.13%	91%	91%	91%
LR	91.24%	90%	88%	89%
VC (LR + SGD)	92.54%	92%	93%	92.05%

5.2. Experiments Using Dataset-2

In addition to dataset-1, this study uses the "clinical trials" dataset to validate the performance of the proposed approach for citation sentiment analysis. Experiments are performed with all the features used for dataset-1 and results are shown in Table 7. Results corroborate the superior performance of the proposed CNN with a 92.25% accuracy when it is used with the combined features. In a similar fashion, the CNN performance is better with fastText and fastText subword, except for GloVe features, where its performance is second to the LSTM model.

Table 7. Experimental results for dataset-2 using all features.

Model	fastText	fastText Subword	GloVe	Combined Features
CNN	89.67%	90.42%	89.81%	92.25%
LSTM	88.88%	89.63%	91.11%	91.89%
RF	89.25%	89.10%	87.62%	91.65%
SGD	86.31%	86.89%	88.65%	89.25%
LR	86.24%	87.21%	87.23%	90.11%
VC (LR + SGD)	88.45%	88.21%	89.65%	91.73%

5.3. Analyzing Importance of Scientific Articles Using Citation Sentiment

In addition to the sentiment classification, the information provided in dataset-1 can be used to analyze the quality of an article by analyzing the number of citations and the ratio of positive, negative, and neutral sentiments regarding that article. For this purpose, the articles with at least 100 citations are selected from dataset-1, and the distribution of the citations is given in Figure 2. Each sub-figure is drawn for one article, and only articles that are cited by 100 or more papers are included. The Figures show the distribution of the citation sentiments of a specific scientific article within other articles, which means the positive, negative, or neutral context in whichever paper is cited. For example, if the paper is praised for its contribution, the sentiment is positive; if it is criticized, the citation sentiment is negative, and so on. This indicates that neutral citations are predominant for research articles, followed by positive citations, while negative citations are the lowest on average. Currently, authors' ranking is determined solely by the number of citations

their works have recieved, which is not correct, as the majority of the citations do not attest to the scientific significance of the articles. Instead, a higher number of citations is neutral. Considering the positive sentiments would be much more appropriate for analyzing the true potential of an article. Similarly, several articles have a higher number of negative citations, say, for example, Figure 2b,f, where the number of positive and negative citations are almost equal; however, both types of citations are considered as positive when determining the importance of those articles. Similarly, Figure 2l shows a higher number of negative citations, yet, negative citations are counted analogously to those that are positive and neutral.

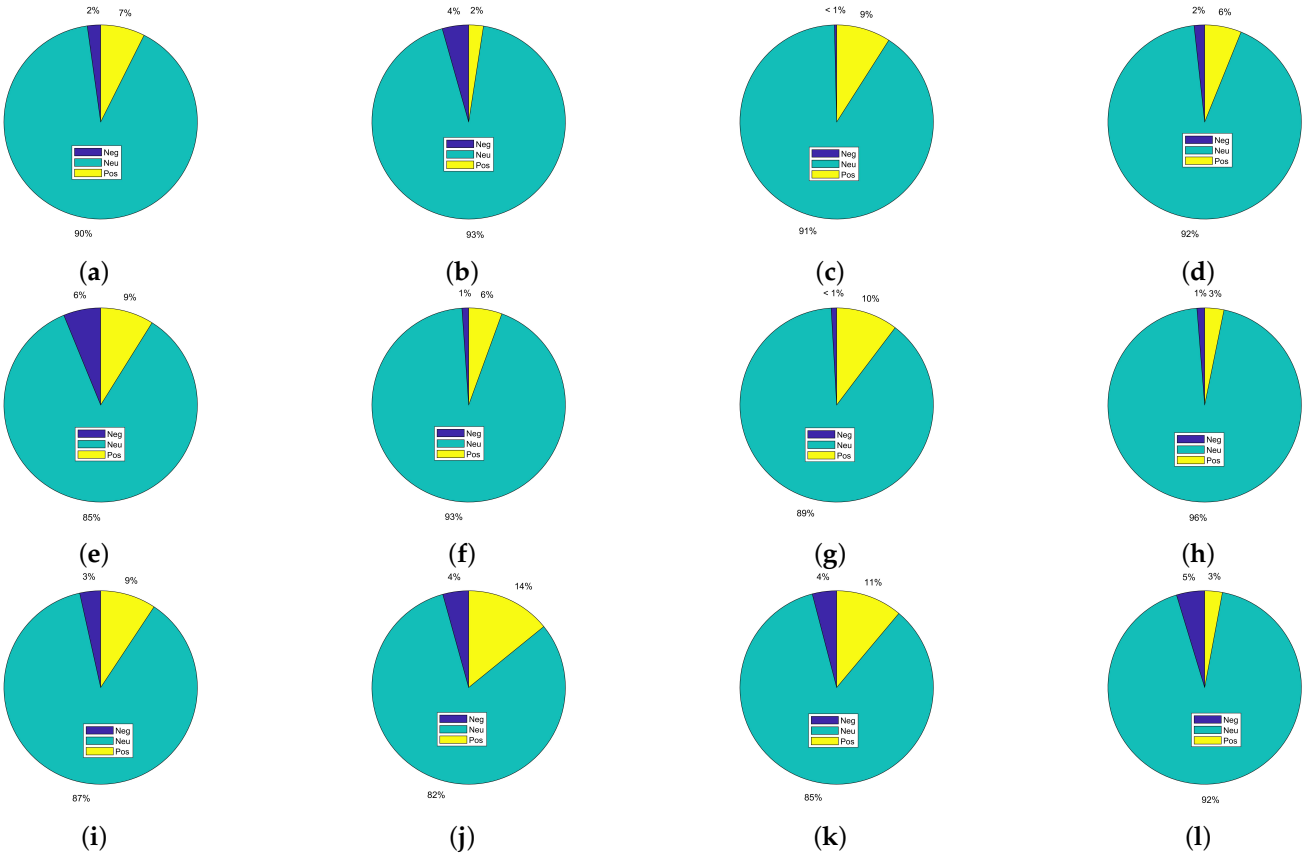

Figure 2. Distribution for negative, neutral, and positive citations from dataset-1 for articles with $\geq$100 citations.

Dataset-1 does not contain the information of the cited article, so the analysis regarding the quality of the article can not be carried out further. The role of sentiment citation can be very influential for determining the true importance of a scientific article, as not every citation is made to acknowledge the superiority of a research article. However, for this purpose, a specialized dataset is needed that contains the citation text, citation sentiment, and the information of the citing and the cited article. For this purpose, we additionally collected a dataset of several articles that have a high number of citations. For this purpose, we selected four articles with citations higher than 1000, as follows:

1. *"New Avenues in Opinion Mining and Sentiment Analysis"*;
2. *"Recent Trends in Deep Learning Based Natural Language Processing"*;
3. *"Techniques and applications for sentiment analysis"*;
4. *"Sentiment analysis algorithms and applications: A survey"*.

The number of citations for these articles is 1316, 2138, 1469, and 2380, respectively. The text where these papers are cited is extracted manually from the papers using Google Scholar, as it provides details of all papers citing a specific paper. The in-text citation senti-

ment extracted from these papers is labeled manually. The distribution of the sentiments of the dataset is shown in Figure 3. On average, neutral sentiments are predominant, followed by positive sentiments, while the ratio of negative sentiments is the lowest. Considering the fact found in dataset-1 and the self-collected dataset, that a higher number of citations has neutral sentiments and does not affirm the importance of an article, it is inappropriate to simply use the number of citations as an indicator for the importance of an article or an author.

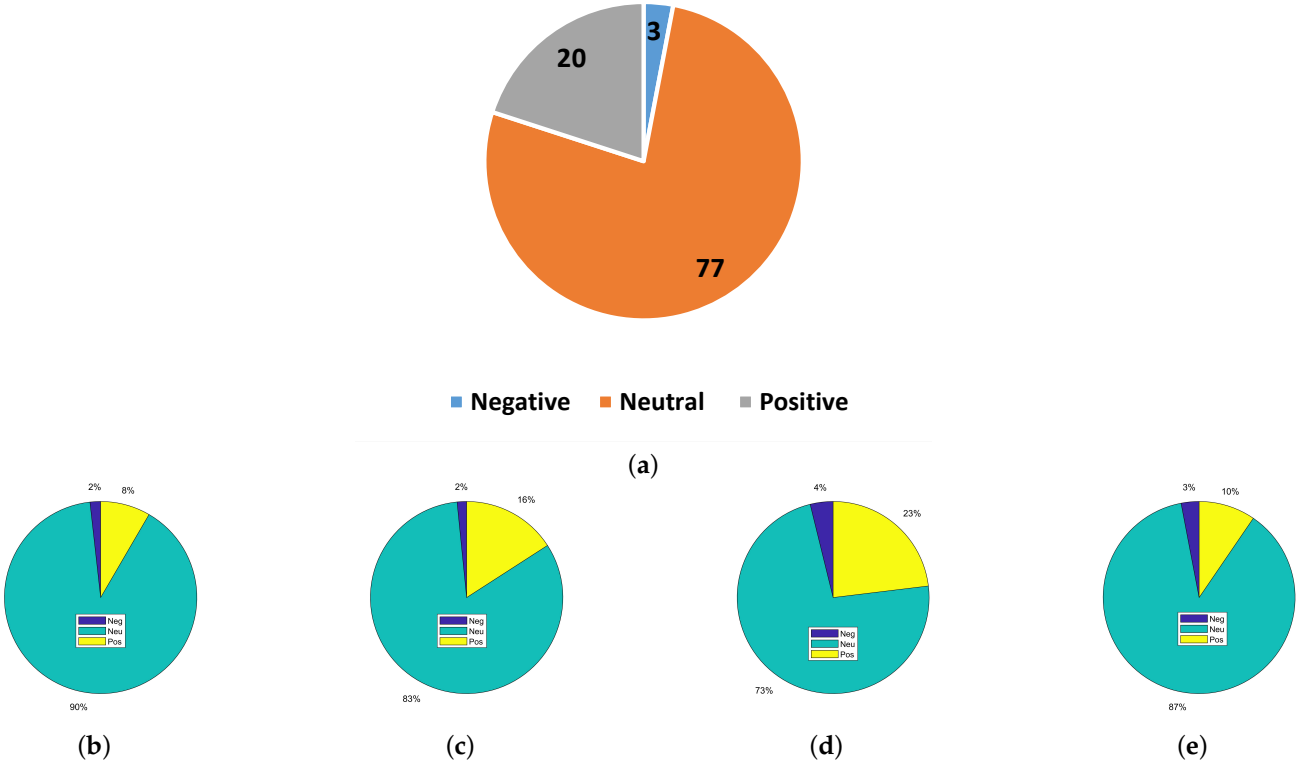

Figure 3. Sentiment distribution of the collected dataset.

5.4. Performance Comparison with State-of-the-Art Studies

A performance comparison is also made with other recent studies to show the significance of the proposed approach. For this purpose, two recent studies have been selected including [12,17]. The study uses several models for citation sentiment analysis, such as support vector classifier (SVM), multinomial naïve Bayes, k-nearest neighbor, and LR. On the other hand, Ref. [12] has been selected, as it uses an ensemble of CNN and LSTM deep learning models. Comparison results provided in Table 8 suggest that the proposed framework with the CNN model using blended features shows better results than the state-of-the-art models for the sentiment analysis of in-text citations.

Table 8. Performance comparison with other approaches.

Reference	Model	Accuracy
[17]	SVC	0.85
[17]	Multinomial NB	0.87
[17]	KNN	0.79
[17]	Logistic regression	0.86
[12]	CNN+LSTM	0.85
Proposed	CNN	0.93

6. Discussions

The state-of-the-art machine learning and deep learning models coupled with word embedding techniques have been explored for the classification of citations into positive, negative, and objective (neutral). Each word embedding technique has been evaluated using standard evaluation measures, such as accuracy, precision, recall, and F1 score. Figure 4 presents the performance comparison of classifiers for sentiment analysis. Results reveal that the deep learning model, CNN, outperforms when trained on blended features.

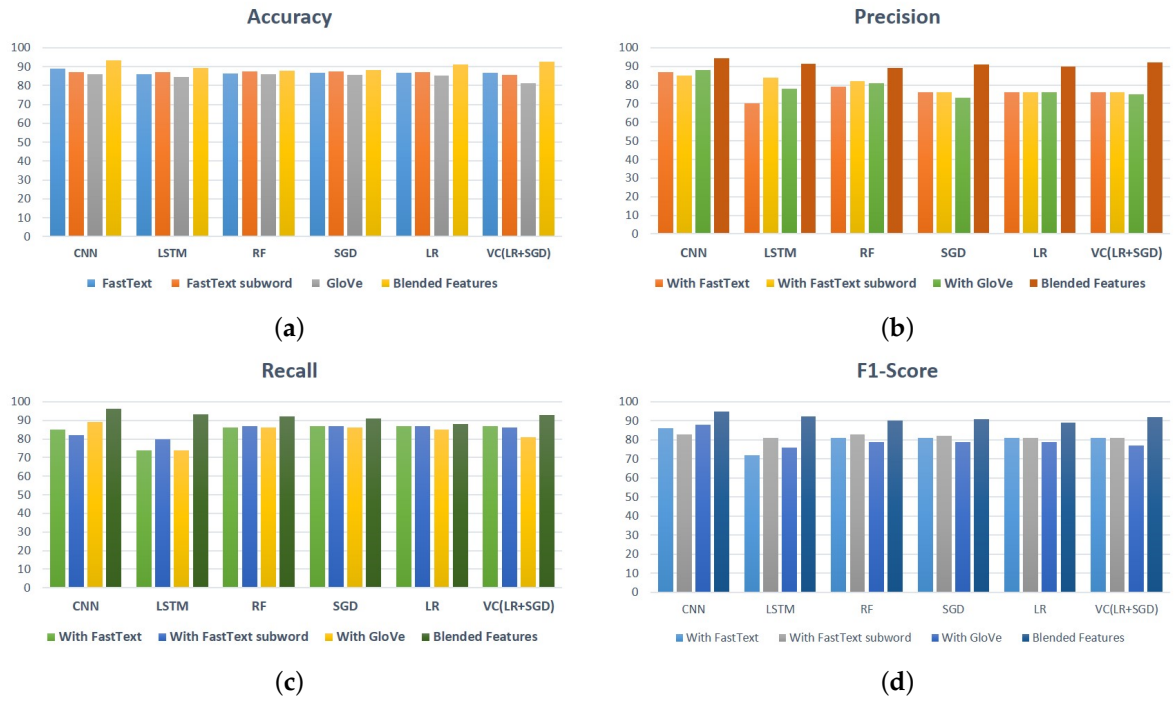

Figure 4. Performance comparison of machine and deep learning models with different features, (**a**) accuracy, (**b**) precision, (**c**) recall, and (**d**) F1 score.

Figure 4a shows the accuracy comparison of models using fastText, fastText subword, and GloVe. It can be observed that CNN achieved the best results with every word embedding technique, as compared to the other models. On the other hand, LSTM has shown the highest accuracy result with combined features and the lowest accuracy with GloVe. Machine learning models and the voting classifier has also shown moderate performance results when trained on fastText. RF, LR, and SGD showed improved results when trained on fastText subword, and the lowest results when trained on GloVe.

Figure 4b presents the performance comparison of models in terms of precision. As demonstrated in the results, CNN achieved the highest values of precision with each word embedding technique. The highest precision achieved by CNN is 94.24% when it is trained on GloVe. LSTM has shown the lowest value for precision, with 70% when trained on fastText, whereas LSTM achieved 84% precision with fastText subword. RF achieved the highest precision score of 81% with GloVe. SGD and LR showed similar values for precision with fastText and fastText subword.

Figure 4c shows the recall comparison of models using fastText, fastText subword, GloVe embedding, and combined features. SGD, LR, and the voting classifier achieved the highest recall, of 96.18%, when used with combined features. Machine learning models achieved the highest score for recall when trained on fastText subword, with 87%. However, CNN achieved the highest recall of 89% when trained on GloVe. LSTM achieved a good score for recall with fastText subword, and achieved the lowest recall score with fastText and GloVe.

Figure 4d manifests the F1 score comparison of models for citation sentiment analysis. Results reveal that the highest F1 score is obtained by CNN with 95% when used with the combined feature. For fastText and fastText subword, the F1 scores are 86% and 83%, respectively. LSTM achieved the lowest F1 score with fastText and GloVe. Machine learning models have shown almost similar results in terms of F1 score for the sentiment analysis of in-text citations.

GloVe has around 840 billion crawl words, which is relatively higher than fastText and fastText subword embedding. However, GloVe also considers global stats about word occurrence and shows better performance as compared to word2vec. However, GloVe did not perform well on citation sentiment analysis, as it only considers word co-occurrence. As a citation text is from multiple domains, fastText improves the result by considering parts of words by using n-grams. This ability makes fastText more generalized for unknown words. The above discussion illustrates that CNN performs better with word embedding as compared to other models. We choose CNN coupled with blended features and PCA as the best approach to perform sentiment analysis on citation texts. The dataset used to train models is imbalanced, with less negative and positive instances than those in the neutral class. CNN has convolutional layers that extract useful features from the text and can find the relationship among features. Word embedding represents text in the form of vectors. In this study, we explored the suitable word combination of classifiers and word embedding techniques to classify imbalanced datasets of citation texts.

7. Conclusions

Keeping in view the fact that all citations are not equally important, and that qualitative measures of frequency count is not appropriate to portray the importance of research work, sentiment analysis of the citation can provide a better insight into the quality of a cited work, as it can predict the positive, negative, and objective intent of the citing author. The quality of a work can be properly judged by dividing the citations into appreciating and criticizing citations. In this regard, this study proposed an approach for the sentiment analysis of in-text citations by exploring multiple word embedding techniques, including fastText, fastText subword, GloVe, and their combination, coupled with machine learning and deep learning models. Both the machine learning and deep learning models are extensively studied, in addition to a customized CNN model. Experimental evaluation indicates that CNN is the best-performing model when coupled with combined features and PCA for citation sentiment analysis. Furthermore, CNN surpassed other models with all three embedding techniques. On average, the proposed CNN, when used with combined features, achieves 93.47% accuracy, 94.24% precision, 96.18% recall, and 95% F1 score.

Future Directions

We outline the following future directions regarding this study. We intend to utilize a deep neural network-based ensemble model with the fusion of word embedding techniques to analyze citation sentiments in the future. The currently available corpus has several shortcomings, and several aspects remain unexplored. Thus, we are planning to develop a corpus annotated with the sentiments according to the context of the citation. Currently, the citations are distributed without considering the type of the citing article, such as journal paper, conference paper, technical report, etc. We intend to analyze the importance of an article by considering the distribution of the citing articles with respect to its type of publication. Future experiments may improve the results further.

Author Contributions: Conceptualization, M.K. and M.M.S.M.; Formal analysis, M.K., M.M.S.M. and M.U.; Funding acquisition, A.M.; Investigation, M.U. and I.A.; Methodology, M.U.; Project administration, M.U., S.S. and M.U.; Resources, A.M.; Software, S.S.; Supervision, I.A.; Validation, M.U. and A.M.; Visualization, S.S.; Writing—original draft, M.K. and M.M.S.M.; Writing—review and editing, I.A. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Peroni, S.; Dutton, A.; Gray, T.; Shotton, D. Setting our bibliographic references free: Towards open citation data. *J. Doc.* **2015**, *71*, 253–277. [CrossRef]
2. Case, D.O.; Higgins, G.M. How can we investigate citation behavior? A study of reasons for citing literature in communication. *J. Am. Soc. Inf. Sci.* **2000**, *51*, 635–645. [CrossRef]
3. Waltman, L. A review of the literature on citation impact indicators. *J. Inf.* **2016**, *10*, 365–391. [CrossRef]
4. Zhu, X.; Turney, P.; Lemire, D.; Vellino, A. Measuring academic influence: Not all citations are equal. *J. Assoc. Inf. Sci. Technol.* **2015**, *66*, 408–427. [CrossRef]
5. Wagner, C.S.; Roessner, J.D.; Bobb, K.; Klein, J.T.; Boyack, K.W.; Keyton, J.; Rafols, I.; Börner, K. Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *J. Inf.* **2011**, *5*, 14–26. [CrossRef]
6. Huggett, S. Journal bibliometrics indicators and citation ethics: A discussion of current issues. *Atherosclerosis* **2013**, *230*, 275–277. [CrossRef]
7. Bonzi, S. Characteristics of a literature as predictors of relatedness between cited and citing works. *J. Am. Soc. Inf. Sci.* **1982**, *33*, 208–216. [CrossRef]
8. Aljuaid, H.; Iftikhar, R.; Ahmad, S.; Asif, M.; Afzal, M.T. Important citation identification using sentiment analysis of in-text citations. *Telemat. Inform.* **2021**, *56*, 101492. [CrossRef]
9. Athar, A. Sentiment analysis of citations using sentence structure-based features. In Proceedings of the ACL 2011 Student Session, Portland, OR, USA, 19–24 June 2011; pp. 81–87.
10. Shahid, A.; Afzal, M.; Qadir, M. Discovering semantic relatedness between scientific articles through citation frequency. *Aust. J. Basic Appl. Sci.* **2011**, *5*, 1599–1604.
11. Mifrah, S.; Hourrane, O.; El Habib Benlahmar, N.B.; Rachdi, M. Citation Sentiment Analysis: A Brief Comprehensive Study. *J. Islam. Ctries. Soc. Stat. Sci.* **2017**, *3*, 145–156.
12. Muppidi, S.; Gorripati, S.K.; Kishore, B. An approach for bibliographic citation sentiment analysis using deep learning. *Int. J.-Knowl.-Based Intell. Eng. Syst.* **2020**, *24*, 353–362. [CrossRef]
13. Ikram, M.T.; Afzal, M.T. Aspect based citation sentiment analysis using linguistic patterns for better comprehension of scientific knowledge. *Scientometrics* **2019**, *119*, 73–95. [CrossRef]
14. Demšar, U.; Harris, P.; Brunsdon, C.; Fotheringham, A.S.; McLoone, S. Principal component analysis on spatial data: An overview. *Ann. Assoc. Am. Geogr.* **2013**, *103*, 106–128. [CrossRef]
15. Yousif, A.; Niu, Z.; Tarus, J.K.; Ahmad, A. A survey on sentiment analysis of scientific citations. *Artif. Intell. Rev.* **2019**, *52*, 1805–1838. [CrossRef]
16. Xu, J.; Zhang, Y.; Wu, Y.; Wang, J.; Dong, X.; Xu, H. Citation sentiment analysis in clinical trial papers. In Proceedings of the AMIA annual Symposium Proceedings, American Medical Informatics Association, San Francisco, CA, USA, 14–18 November 2015; Volume 2015, p. 1334.
17. Amjad, Z.; Ihsan, I. VerbNet based citation sentiment class assignment using machine learning. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 621–627. [CrossRef]
18. Parthasarathy, G.; Tomar, D. Sentiment analyzer: Analysis of journal citations from citation databases. In Proceedings of the 2014 5th International Conference-Confluence the Next Generation Information Technology Summit (Confluence), Noida, India, 25–26 September 2014; IEEE: Piscataway, NJ, USA, 2014, pp. 923–928.
19. Parthasarathy, G.; Tomar, D. A survey of sentiment analysis for journal citation. *Indian J. Sci. Technol.* **2015**, *8*, 1–8. [CrossRef]
20. Nazir, S.; Asif, M.; Ahmad, S. Important Citation Identification by Exploiting the Optimal In-text Citation Frequency. In Proceedings of the 2020 International Conference on Engineering and Emerging Technologies (ICEET), Lahore, Pakistan, 22–23 February 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6.
21. Umer, M.; Sadiq, S.; Missen, M.M.S.; Hameed, Z.; Aslam, Z.; Siddique, M.A.; Nappi, M. Scientific papers citation analysis using textual features and SMOTE resampling techniques. *Pattern Recognit. Lett.* **2021**, *150*, 250–257. [CrossRef]
22. Raza, H.; Faizan, M.; Hamza, A.; Mushtaq, A.; Akhtar, N. Scientific text sentiment analysis using machine learning techniques. *Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*, 157–165. [CrossRef]
23. Joulin, A.; Grave, E.; Bojanowski, P.; Douze, M.; Jégou, H.; Mikolov, T. Fasttext. zip: Compressing text classification models. *arXiv* **2016**, arXiv:1612.03651.
24. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]

25. Pennington, J.; Socher, R.; Manning, C.D. GloVe: Global Vectors for Word Representation. In Proceedings of the Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
26. Albawi, S.; Mohammed, T.A.; Al-Zawi, S. Understanding of a convolutional neural network. In Proceedings of the 2017 International Conference on Engineering and Technology (ICET), Antalya, Turkey, 21–23 August 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–6.
27. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. *Neural Inf. Process. Syst.* **2012**, *25*, 84–94. [CrossRef]
28. Castiglione, A.; Vijayakumar, P.; Nappi, M.; Sadiq, S.; Umer, M. COVID-19: Automatic Detection of the Novel Coronavirus Disease from CT Images Using an Optimized Convolutional Neural Network. *IEEE Trans. Ind. Informatics* **2021**, *17*, 6480–6488. [CrossRef]
29. Li, Q.; Cai, W.; Wang, X.; Zhou, Y.; Feng, D.D.F.; Chen, M. Medical image classification with convolutional neural network. In Proceedings of the 2014 13th International Conference on Control Automation Robotics and Vision, ICARCV 2014, Singapore, 10–12 December 2014; pp. 844–848. [CrossRef]
30. Pereira, S.; Pinto, A.; Alves, V.; Silva, C. Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images. *IEEE Trans. Med Imaging* **2016**, *35*, 1. [CrossRef] [PubMed]
31. Bar, Y.; Diamant, I.; Wolf, L.; Lieberman, S.; Konen, E.; Greenspan, H. Chest pathology detection using deep learning with non-medical training. *Proc.-Int. Symp. Biomed. Imaging* **2015**, *2015*, 294–297. [CrossRef]
32. Luo, L.x. Network text sentiment analysis method combining LDA text representation and GRU-CNN. *Pers. Ubiquitous Comput.* **2019**, *23*, 405–412. [CrossRef]
33. Song, S.; Huang, H.; Ruan, T. Abstractive text summarization using LSTM-CNN based deep learning. *Multimed. Tools Appl.* **2019**, *78*, 857–875. [CrossRef]
34. Banerjee, I.; Ling, Y.; Chen, M.C.; Hasan, S.A.; Langlotz, C.P.; Moradzadeh, N.; Chapman, B.; Amrhein, T.; Mong, D.; Rubin, D.L.; et al. Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification. *Artif. Intell. Med.* **2019**, *97*, 79–88. [CrossRef]
35. Dai, L.; Sheng, B.; Wu, Q.; Li, H.; Hou, X.; Jia, W.; Fang, R. Retinal microaneurysm detection using clinical report guided multi-sieving CNN. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Quebec City, QC, Canada, 11–13 September 2017; Springer: Berlin/Heidelberg, Germany, 2017; pp. 525–532.
36. Biau, G.; Scornet, E. A random forest guided tour. *Test* **2016**, *25*, 197–227. [CrossRef]
37. Khalid, M.; Ashraf, I.; Mehmood, A.; Ullah, S.; Ahmad, M.; Choi, G.S. GBSVM: Sentiment classification from unstructured reviews using ensemble classifier. *Appl. Sci.* **2020**, *10*, 2788. [CrossRef]
38. Umer, M.; Ashraf, I.; Mehmood, A.; Ullah, S.; Choi, G.S. Predicting numeric ratings for Google apps using text features and ensemble learning. *Etri J.* **2021**, *43*, 95–108. [CrossRef]
39. Wright, R.E. Logistic regression. In *Reading and Understanding Multivariate Statistics*; American Psychological Association: Washington, DC, USA, 1995; pp. 217–244.
40. Gardner, W.A. Learning characteristics of stochastic-gradient-descent algorithms: A general study, analysis, and critique. *Signal Process.* **1984**, *6*, 113–133. [CrossRef]