Journal of Ambient Intelligence and Humanized Computing (2024) 15:419–433
https://doi.org/10.1007/s12652-022-03899-6

ORIGINAL RESEARCH

A deep learning approach for context-aware citation recommendation using rhetorical zone classification and similarity to overcome cold-start problem

Muhammad Azeem Abbas[1] · Saheed Ajayi[2] · Muhammad Bilal[3] · Ade Oyegoke[2] · Maruf Pasha[4] · Hafiz Tauqeer Ali[5]

Received: 1 May 2021 / Accepted: 4 May 2022 / Published online: 28 May 2022

Abstract
In the recent decade, the citation recommendation has emerged as an important research topic due to its need for the huge size of published scientific work. Among other citation recommendation techniques, the widely used content-based filtering (CBF) exploits research articles' textual content to produce recommendations. However, CBF techniques are prone to the well-known cold-start problem. On the other hand, deep learning has shown its effectiveness in understanding the semantics of the text. The present paper proposes a citation recommendation system using deep learning models to classify rhetorical zones of the research articles and compute similarity using rhetorical zone embeddings that overcome the cold-start problem. Rhetorical zones are the predefined linguistic categories having some common characteristics about the text. A deep learning model is trained using ART and CORE datasets with an accuracy of 76 per cent. The final ranked lists of the recommendations have an average of 0.704 normalized discounted cumulative gain (nDCG) score involving ten domain experts. The proposed system is applicable for both local and global context-aware recommendations.

Keywords Content-based filtering · Cold-start · Bi-LSTM

1 Introduction

The ever-increasing size of digital libraries confronts researchers with the problem of information overload (Mahdi et al. 2020). Finding the most relevant research articles is still challenging for the researchers, especially when exploring any new research domain. To address the problem, the present work proposes a citation recommendation system to help researchers find relevant articles from the huge and complex landscape. Recommender systems are based on three main models, namely: content-based filtering (CBF), collaborative filtering (CF) and graph-based (GB) have emerged as a solution for finding similar articles (Ma et al. 2020). Collaborative filtering-based approaches employ information from user profiles such as past interactions, feedback or ratings and friends network to make recommendations about papers (Martins et al. 2020). CF suppose that users with a common interest will like similar items. Several variants of the CF-based approaches were proposed to improve the accuracy of the provided recommendations (Wang et al. 2020b). The quality of CF-based approaches is highly dependent on available user rating or feedback information. Unavailability or partial availability of such information leads to a sparsity problem due to many missing values in the user-paper matrix (Ali et al. 2020a). Moreover, collaborative filtering hardly addresses the recommendations for a new research problem. Similarly, graph-based models (GB) represents articles/citation information as nodes and edges connected to form a network (Wang et al. 2020a). Article recommendations are made through graph traversal or the link prediction method. Graph-based approaches inhabit the problem of over-weighting where old and outdated articles remain the same in the network.

✉ Muhammad Azeem Abbas
m.abbas@uwtsd.ac.uk

1 University of Wales Trinity Saint David, Birmingham Campus, Birmingham, UK

2 School of Built Environment, Engineering and Computing, Leeds Beckett University, Leeds, UK

3 Big Data Enterprise and Artificial Intelligence Laboratory (Big-DEAL), Bristol Business School, University of West of the England, Bristol, UK

4 Bahauddin Zakariya University, Multan, Pakistan

5 Taif University, Taif, Saudi Arabia

However, the new articles are not linked to the network because they do not have a direct link with the existing nodes. In contrast, CBF approaches exploit the content of the research article to produce recommendations (Habib and Afzal 2019). CBF only performs well when user preferences and article descriptions are provided; otherwise, such techniques are prone to well-known cold-start problem (Ali et al. 2020a; Martins et al. 2020). Cold-start problem concerns that the system cannot draw any conclusion for articles about which it has not yet gathered enough information.

The present work addresses the cold-start problem by computing content-based similarity among articles, even if user preferences and article descriptions are not provided. The traditional coarse-grained similarity computation does not consider multi-facets or more concisely semantic facets that reflect actual similarity among documents. The current proposal follows the argument made by Bär et al. (2011) that items are similar if a given facet of similarity relates. In the case of research articles, the similarity would be the multi-facets of research, e.g., goal, methodology, findings, results and conclusion. Linguistically, these facets are called rhetoric zones. The present work provides tailored recommendations according to the rhetoric zones. For example, a recommendation shall be made with a similar problem statement but different methods or similar methodology but different findings.

The proposal presented here utilizes the deep learning (a part of machine learning) method to classify rhetoric zones of the research articles and compute zone-wise similarities to get a ranked list of relevant citations (shown in Fig. 1). Formally, a query article (D_q) and set of articles (D) containing both relevant and irrelevant documents with respect to D_q are provided. All these articles containing a set of sentences (S_1,S_2,..) are transformed to their rhetoric zones representation ($D_i^{RZ} \Big| i = 1 \; to \; n \; \& \; D_q^{RZ}$). This transformation is carried out by the proposed deep learning model. Finally, the goal is to retrieve a ranked list of articles based on similarity scores between the query and candidate articles.

Given $D = \{D_1, D_2, D_3, \ldots, D_n\}$ & D_q

$$D_1^{RZ} = \{ (S_1, S_2, \ldots, S_w)^{BACK}, (S_1, S_2, \ldots, S_x)^{MOTIV}, (S_1, S_2, \ldots, S_y)^{PROB}, (S_1, S_2, \ldots, S_z)^{GOAL}, \ldots \}$$

$$D_2^{RZ} = \{ (S_1, S_2, \ldots, S_w)^{BACK}, (S_1, S_2, \ldots, S_x)^{MOTIV}, (S_1, S_2, \ldots, S_y)^{PROB}, (S_1, S_2, \ldots, S_z)^{GOAL}, \ldots \}$$

$\ldots$

$$D_n^{RZ} = \{ (S_1, S_2, \ldots, S_w)^{BACK}, (S_1, S_2, \ldots, S_x)^{MOTIV}, (S_1, S_2, \ldots, S_y)^{PROB}, (S_1, S_2, \ldots, S_z)^{GOAL}, \ldots \}$$

$$D_q^{RZ} = \{ (S_1, S_2, .., S_{w\}})^{BACK}, (S_1, S_2, .., S_{x\}})^{MOTIV}, (S_1, S_2, .., S_{y\}})^{PROB}, (S_1, S_2, .., S_{z\}})^{GOAL}, \ldots \}$$

$$ListofArticles \leftarrow Rank(Similarity \left(D_q^{RZ}, (D_1^{RZ}, D_2^{RZ}, \ldots, D_n^{RZ}) \right))$$

Recently, deep learning methods for research papers recommendations have shown significant improvements due to their ability to capture the contextual information and semantic representations of the facets of the research articles (Bai et al. 2019; Bansal et al. 2016; Zeng

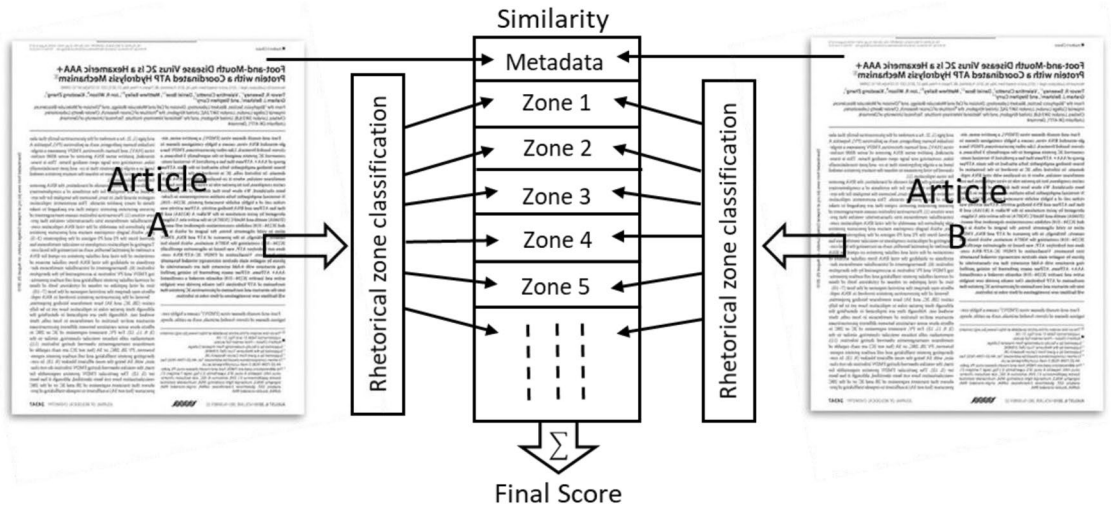

Fig. 1 Overview of the rhetoric zone classification and similarity

and Acuna 2020). However, very few researchers using deep learning have addressed the problem of cold-start and nearly all the research conducted has only focused on personalized recommendations (Ali et al. 2020a). Recommendations are made using the user's profile information and history in personalized recommendation models. In contrast, non-personalized recommendation model generates uniform recommendations for all users containing relevant and top-rated articles. The proposed technique provides a solution for the cold-start problem and generates non-personalized recommendations by computing rhetoric zones similarity for both article-to-article and article-to-user query. A research article contains several rhetoric zones with specific characteristics (Asadi et al. 2019; Badie et al. 2018). These rhetoric zones can be classified as background, motivation, goal, problem, hypothesis, method, model, experiment, results/findings and conclusion (Liakata and Soldatova 2009; Liakata et al. 2009). The proposed rhetoric zone classification method retrieves small chunks of text from the published articles against the above rhetorical zones and computes zone-wise similarity. In addition to rhetoric zones similarity, the traditional metadata comparison is in-cooperated to evaluate its effectiveness. The deep learning model is trained and tested on CORE (Knoth et al. 2017) and ART (Liakata and Soldatova 2009) datasets. The tenfold crossover validation of the trained model has resulted in an accuracy of 76.3%. The recommendations made by the proposed technique are evaluated against 2543 articles using average precision and normalized discounted cumulative gain (nDCG) measures involving ten domain experts.

The present work makes a noteworthy contribution to the cold-start problem in citation recommendation through rhetoric zones classification using a deep learning model and computing similarity among rhetoric zones. Moreover, the present work performs an extensive evaluation for measuring the effectiveness of the proposed model using a combination of two real-world datasets. The paper is organized as follows: Sect. 2 contains a literature review of specifically, deep learning-based research article recommendation techniques. In Sect. 3, the proposed system's methodological details are provided, explaining the process of rhetoric zones classification and similarity computation. Section 4 presents the experimental results. Finally, Sect. 5 concludes the present work with future directions.

2 Related work

The traditional citation recommendation approaches are based on co-citations, bibliographic coupling, metadata analysis, content-based filtering, collaborative filtering and graph-based filtering (Habib and Afzal 2019). Co-citations or direct citation or the bibliographic coupling recommend research articles based on citation analysis, mainly the relationship information among articles. Any research article is considered relevant to another article if a citation link is present between them. The citation link can be a direct link from one article to another, or it can be through some intermediate article. The problem with the co-citations and bibliographic coupling is that the recommendation works only on the explicit information provided as citation links. Articles that are relevant but not cited directly or indirectly will not appear in citation-based recommendation techniques. The citation-based techniques' accuracy was enhanced by combining the content of the cited articles and the citation information. However, content-based filtering (CBF) methods have their own challenges (Ma et al. 2020).

Content-based filtering (CBF) approaches exploits the content of the research article to produce recommendations. CBF approaches utilize the article's title, keywords, abstract, venue and authors information, and in some cases, the whole article itself. However, the complete article reduces the recommendation accuracy as the article contains a substantial number of wider or general context statements. Suppose the problem addressed by an article is the same as query paper, but the methodology of the article is entirely different from query paper; this results in a very weak similarity between the article and the query paper if the content of the complete article is considered. The reason is that a large portion of the article content is about its methodology with significantly overlapping/similar content with the query paper. On the other hand, keywords were the first choice of CBF approaches where keywords extracted from the articles were matched to compute similarity. Several approaches enhance extracted keywords by augmenting them using dictionaries or ontologies (Chughtai et al. 2020). However, the keyword-based search is a straightforward technique but with several limitations such as keywords does not reflect the whole article and the user needs, ambiguous keywords and vocabulary mismatched.

CBF only perform well when user preferences along with article descriptions are provided; otherwise, such techniques are prone to the well-known cold-start problem. The cold-start problem concerns the issue that the system cannot draw any conclusion for articles about which it has not yet gathered enough information. This problem is highly observed for newly published articles as they are not cited by many papers and their collaborative ratings are also unavailable (Abro et al. 2020; Christoforidis et al. 2018).

Several researchers have recently employed deep learning models for citation recommendations. These deep learning citation recommendation approaches have used paper's content, profile information, keywords, and venue information to train the deep learning models, which later makes recommendations (Ambalavanan and Devarakonda 2020;

Jeong et al. 2020; Kumar et al. 2021). Deep learning-based approaches have shown better results as compared to matrix-based and graph-based citation recommendation techniques. However, very few have addressed the cold-start problem and mainly the global recommendation context. Global and local are two types of context-aware citation recommendations (Jeong et al. 2020; Wang et al. 2020a). Global context-aware citation recommendation techniques consider the title and abstract of the query paper and candidate citation paper to deriving the recommendations. In the local context, the text nearby a citation reference is considered for providing recommendations.

A recommender system named HRM (Li et al. 2019) was proposed, which sends out newsletters containing citation recommendations to the subscribed users. Citation recommendations are generated based on the user's browsing history (previous search queries and interactions) on the articles search engine. The newsletter items are ranked citation recommendations. This approach faced the cold-start problem for new users who do not have browsing history or just have subscribed. HRM system has made recommendations based on entity (authors, articles, venue) similarity in embeddings space. A usability approach of recording user interaction by monitoring clicks made by a user is used to make recommendations for a new user. HRM combines entity information with user behavior to generate the newsletter items list.

Another approach using deep learning was presented to overcome the cold-start problem in a collaborative filtering scenario (Bansal et al. 2016). This technique has used gated recurrent units (GRU) to train text sequences for collaborative filtering tasks. However, collaborative filtering approaches are prone to sparsity problem where data about user interaction is unavailable. Bansal et al. has combined metadata of the article with collaborative information (graph structure) to generate first recommendations for a user. A graph representation from the author and article profile was constructed in heterogeneous information networks for citation recommendation by Ma and Wang (2019) in a system named HGRec. HGRec initializes the node vector by using word-embeddings of the text extracted from the candidate articles. Later, graph representation is updated by joining it with node embeddings using a meta-path based proximity measure. Like HRM and Bansal et al. (2016), HGRec uses embeddings for similarity computation. A system named HIPRec (Xiao Ma et al. 2019) for citation recommendation claimed that previous techniques had computed similarity from a bipartite network of query and candidate article. However, other networks information such as venue, researchers, topic, research domains have been included in HIPRec to form a meta-graph that increases the accuracy of the recommendations. A greedy approach is employed to extract sub-graphs for final recommendations. HIPRec is implemented using the DBLP dataset that is mainly a citation graph rather than full-text articles.

In conclusion, nearly all the studies addressing the cold-start problem are based on collaborative filtering, especially those using deep learning techniques. Studies using deep learning have utilized embeddings generated from the deep learning models to compute similarities among modelled information, either items, authors, venues or topics. The traditional multi-layered perceptron (MLP), support vector machine (SVM) and logistics models (Asadi et al. 2019) were the main choices of the previous citation recommendation systems. Auxiliary information such as the author's information, venue, keywords, and social interactions was used in previous citation recommendation systems to overcome the cold-start problem. The same has been reported in recent surveys on deep learning based citation recommendations (Ali et al. 2020a; Martins et al. 2020). In contrast, the present proposal is on content-based filtering and for the reason, only deep learning based hybrid approaches combining collaborative filtering and to some extent the content of the articles to address cold-start problem are presented here. The present work addresses the gap of solving cold-start problem through a content-based filtering approach using deep learning models, which deemed as a novelty to the present research.

3 Rhetorical zone classification and similarity

The architecture of the proposed context-aware citation recommendation system is shown in Fig. 2. The system architecture is comprised of three modules (i) model training, (ii) model testing, and (iii) similarity computation. The training module takes the textual format dataset and generates a trained model for classifying rhetoric zones. The testing module uses the trained model and predicts the class label of new articles. Finally, the similarity phase computes the similarity between rhetorically classified articles and generates a ranked list of articles. Performance evaluation is performed on the accuracy of the trained model and the final ranked list.

3.1 Dataset

The deep learning models are trained on ART (Liakata and Soldatova 2009) and CORE (Knoth et al. 2017) datasets as two well-known corpora. The ART corpus consists of 3433 (Mean 343 Std. Div 163.91) labelled sentences retrieved from 150 research articles of physical chemistry and biochemistry domains. These sentences were taken from the abstract and introduction sections of the articles. Every sentence is manually labelled with a rhetoric zone from a total of ten zones. ART corpus is a small-sized dataset from the

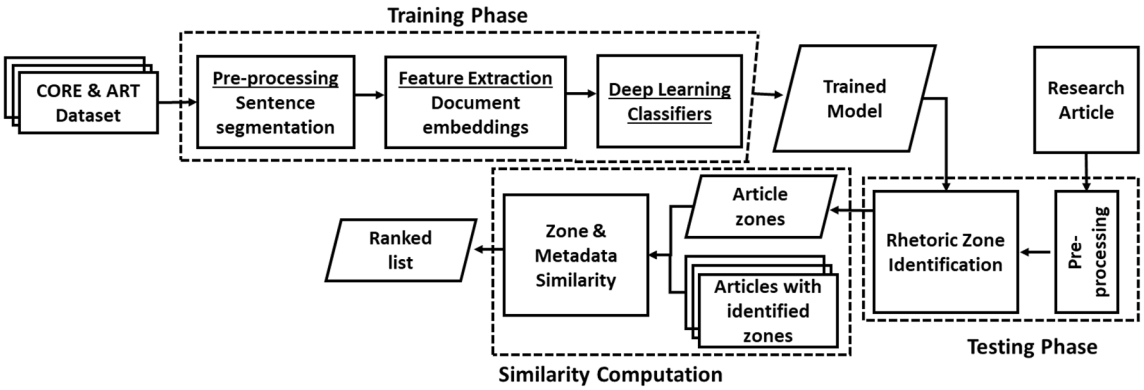

Fig. 2 Proposed system architecture

deep learning perspective. On the other hand, the process of automated data augmentation to increase the size of the dataset is at an early stage of research for the textual data as compared to computer vision. An attempt has been made with online and offline data augmentations following consistency regularization and with the most recent AugLy from Facebook. However, both approaches result in the loss of rhetoric semantics which is salient to the present research. Moreover, the proposed deep learning model was tested with and without functional regularization such as dropout but did not found a significant difference. For this reason, 24,323 sentences were tokenized and extracted from the introduction section of the top 500 open access research articles from the CORE dataset—computer science domain. These sentences were then provided to 60 postgraduate students through an online system for labelling them against ten rhetoric zones (same as ART corpus). The postgraduate students have performed labelling as an ungraded assignment for the research methodology module. A sentence is provided only once to a group of six students. Students can either assign a class label to a sentence or skip it if they reckon the sentence belong to a general category. The labelled sentences are accepted based on the level of agreement among annotators. A total of 18,413 sentences were labelled, among which 13,730 were selected based on their high Cohen's Kappa agreement value, i.e., more than 0.8. After combining the ART with our manually labelled dataset and applying dataset balancing resulted in 14,689 sentences as our final dataset. For the present work, a simple undersampling technique named Neighborhood Cleaning Rule (NCL) which is based on Edited Nearest Neighbor (ENN) method is used for balancing the dataset. The choice of undersampling as compared to oversampling is because the variability is not high (Mean 1716.4 Std. Div 172.70) among class instances, as shown in Fig. 3. Moreover, researchers have reported that simple undersampling outperforms state-of-the-art Synthetic Minority Over-sampling TEchnique (SMOTE) in many cases because SMOTE without variable selection biases the classifiers towards minority classes (Blagus and Lusa 2013). After balancing the dataset, the mean value of the class instances is 1469, with a standard deviation of 9.01.

3.2 Word embedding and feature modelling

Textual data need to be translated into a structure called embeddings (Si et al. 2019) that deep learning algorithms can process it easily and efficiently. For deep learning, the text's vocabulary is a high dimensional vector that can be modelled into low-dimensional, learned continuous vector representation called embeddings. In natural language processing (NLP), word embeddings are used to represent a dense vector of words in low-dimensional space to capture the semantics and syntactic information of the given text (Ali et al. 2020b). Deep learning classifiers can perform mathematical operations on the numerically represented semantics in the word embeddings. Word embeddings support contextual representation, e.g. apple fruit and apple electronics shall be treated differently based on their separate vectors. Some of the well-known models for word embedding are Word2Vec (Mikolov et al. 2013), doc2vec (Han et al. 2018) and most recently BERT (Devlin et al. 2019) by Google. The present work utilizes Word2vec and BERT models for representing the dataset as word embeddings. In addition to word embeddings, the traditional approach of feature extraction using minimum inverse document frequency (min_idf) is performed for comparison purposes. Min_IDF collects the most common and important features from the given text. Furthermore, a feature vector with the size of 2678 was handcrafted by three domain experts using the dataset itself and several other available phrasebooks (Manchester phrasebank, style of writing, etc.) that contains general-purpose rhetoric sentences for technical writing support. This manual feature extraction was initiated due to an initial analysis of the features extracted by the embedding

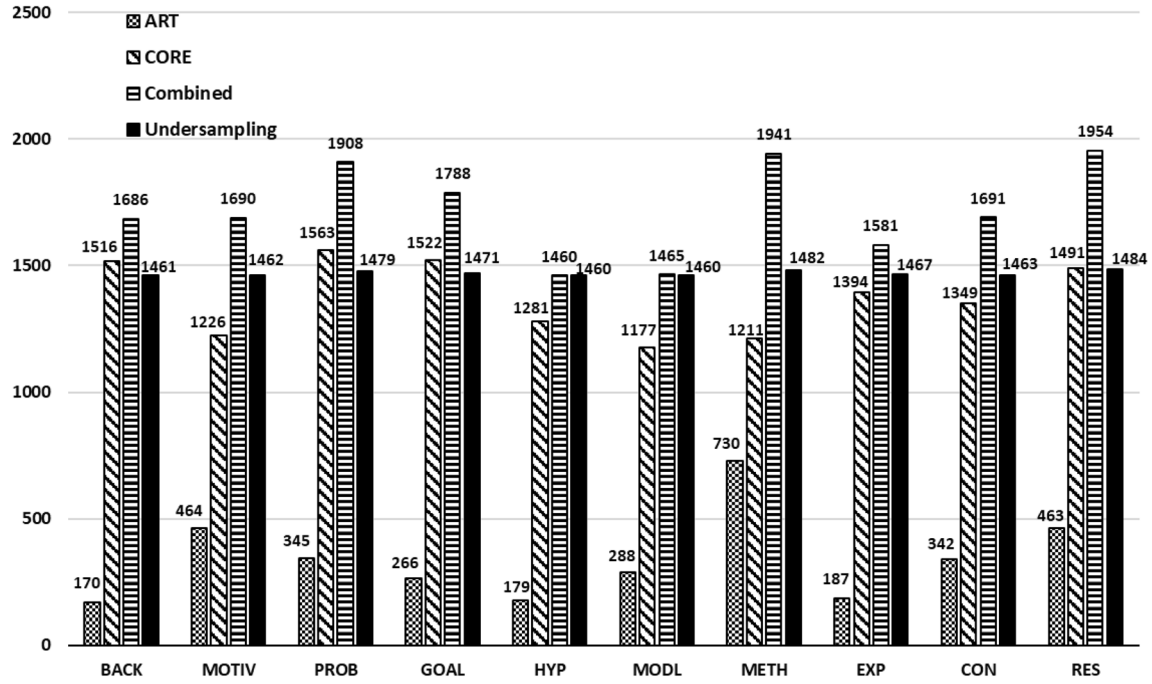

Fig. 3 Dataset class distribution

and Min_IDF methods. Both methods have mainly formulated unigram features, whereas it is assumed that bi-gram or tri-gram features containing stop words might reflect the accurate representation of a sentence.

The traditional feature extraction method lacks the representation of the surrounding context of a word as it merges all possible meanings of the word into a single representation. Word2vec addresses this problem by directly modeling the context of the word in a multidimensional vector representation. This vector representation is the initial task for the predictive models in information and semantic retrieval. The continuous Bag of Words (CBOW) component of the Word2vec infers the target word for a given context, and on the other hand, the skip-gram component infers the context for a given word.

BERT embedding is a recent advancement in modelling the contextual representation of a word or phrase (Ambalavanan and Devarakonda 2020; Jeong et al. 2020). The BERT embeddings can model the contextual information and dynamically modify a multilayer representation, unlike the Word2vec embeddings, which construct a separate vector for each word that remains constant throughout the later processing. The process of learning this contextual information for the construction of embeddings is known as pretraining. After pretraining, sentences are formed with vector representations of the words and fed to classifiers for the prediction. BERT models deeper contextual information as compared to its predecessors due to its underlying deep bi-directional transformer technique. Self-attention transformer architecture is employed by the BERT that provides long-distance context comprehension. BERT fine-tune language model is integrated into the downstream task to achieve task-specific architecture.

3.3 Deep learning classifiers

Long-term short memory (LSTM) is a sequence-based classification method that has shown significant improvements over the traditional text classification methods (Jang et al. 2020; Wang et al. 2020a; Zeng and Acuna 2020). Facebook and Microsoft have claimed over 95% accuracy in automatic translation for their billion size datasets using LSTM. LSTM has improved the recurrent neural network (RNN) architecture by overcoming the vanishing gradient problem by inducing the gating mechanism. Different gates such as input, forget and output decides about retaining the data from the previous state or losing it during the current state. LSTM's ability to extract vital information has shown an important role in text classification. In recent years, the scope of application of LSTMs has rapidly expanded, and several researchers have revamped the LSTM to gain improved accuracy, such as bi-directional long-term short memory (Bi-LSTM).

The Bi-LSTM consists of LSTM units that function in both directions, keeping track of past and future context information. This is done by combining the outputs of two LSTMs layers. The first layer process from backwards to forwards, the other from forwards to backwards. This

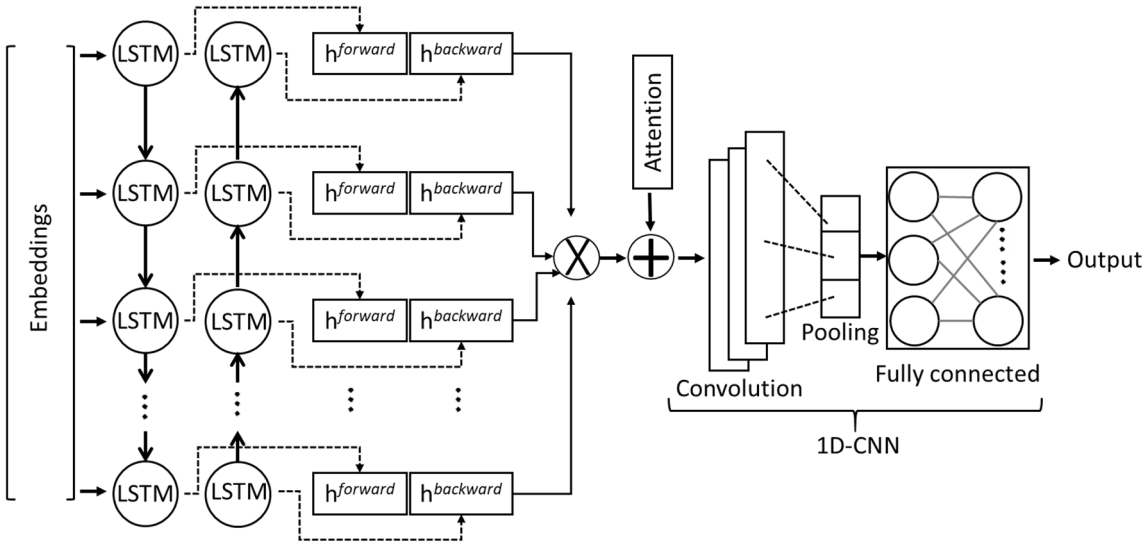

Fig. 4 The Bi-LSTM architecture for rhetoric zone classification

bi-directional approach captures the dependencies between contexts (Fig. 4). Formally, the rhetorical zone as context c_i for the present case is a combination of words $w_i \in \boldsymbol{R}^{d_w}$ that represents a specific semantics. These words in the form of embeddings are inputs and are assembled into matrix $X_i^c \in \boldsymbol{R}^{d_w \times N_i}$. The Bi-LSTM applied over matrix X_i^c is provided as following equations:

$$H_t^{(forward)} = LSTM^{(forward)}(W_t, H_{t-1}^{(forward)})$$

$$H_t^{(backward)} = LSTM^{(backward)}(W_t, H_{t-1}^{(backward)})$$

$$H_t = H_t^{(backward)} \oplus H_t^{(forward)}$$

where $H_t^{(forward)}$ and $H_t^{(backward)}$ represent the hidden states of the forward and backward LSTMs at time t. In Bi-LSTM both backward and forward hidden states are concatenated ($\oplus$) together. The LSTM has a gated approach to overcome the short-memory problem is through adding a cell to justify that either retaining information is utile or not. LSTM cell memory is consisting of input, forgot, and an output gate, i.e., mathematically represented as:

$$Input_t = sigmod(W_{input}x_t + W_{input}h_{t-1} + Bias_{input})$$

$$Forgot_t = sigmod(W_{forgot}x_t + W_{forgot}h_{t-1} + Bias_{forgot})$$

$$Gate_t = tanh(W_{gate}x_t + W_{gate}h_{t-1} + Bias_{gate})$$

$$Output_t = sigmod(W_{output}x_t + W_{output}h_{t-1} + Bias_{output})$$

$$Forgot_t \otimes State_{t-1} + Input_t \otimes Gate_{t-1}$$

$$Hidden_t = Output_t \otimes \tanh\left(State_t\right).$$

where W is the parameters, x_t is the input at time t. The hidden state at time t is computed by the dot product ($\otimes$) of the output gate and tangent activation function over LSTM cell state ($State_t$). The input, forgot, output and cell gates control the information that needs to be retained or passed to the next step. Bi-LSTM can be combined with attention technique to make predictions more precise. Although, after bi-directional LSTM the input word embeddings are shrunk enough to make a prediction using a classifier. However, the correlation between an individual rhetoric zone and the research domain of the article is not clearly visible to the classifier. For this reason, the attention layer with meta-data embeddings are concatenated with the features extracted and a one-dimensional convolutional neural network (1D-CNN) is applied for the final classification task. The attention is given as a maximizing function:

$$\log P(z|X^d, M^d) = \sum_i^k \log P\left(z_i \middle| z_{\leq i}, s\right)$$

where $P\left(z_i \middle| z_{\leq i}, s\right) = softmax(Vh_i)$

$P\left(z_i \middle| z_{\leq i}, s\right)$ is the conditional probability of all previous words in the rhetoric sentences prior to the i-th word. The X^d denotes the vector representation of the rhetoric sentences and k is the number of words in a given rhetoric sentence. The attention is ranked based on metadata context vector representation M^d containing title, keywords, venue and authors information.

In addition to LSTM and Bi-LSTM, the present work has evaluated SciBERT, a large-scale pre-trained model based on BERT. SciBERT follows the same multi-layered bidirectional transformer model as BERT; however, the difference is that SciBERT is pretrained on scientific articles dataset. SciBERT is an uncased BERT model trained on random a sample of over 1.4 million scientific articles from the Semantic Scholar dataset. The pretraining carried out for SciBERT was unsupervised on a multi-domain corpus of scientific articles for improving the performance of NLP tasks such as sentence classification, sequence tagging and dependency parsing. The dataset used by SciBERT closely resembles the present work as it consisted of 18% of research articles from the computer science domain and 82% from the biomedical domain. SciBERT has a 46% vocabulary overlapping with BERT with a total of 3.17 billion tokens.

LSTM and Bi-LSTM models are trained using Adam (Adaptive Moment Estimation) optimization algorithm. Adam optimizer is based on RMSProp (Root Mean Square Propagation) in which learning rate is adapted for each parameter; however, the present work used a fixed learning rate of 0.01. Leaning rate is used to tune parameters in an optimization algorithm to decide step size for reaching a minimum of the loss function. These models were trained with a SoftMax activation function with a batch size of 128 and L2 regularization. The embeddings generated by Word-2vec and BERT was lowercased unigrams vocabulary tokens of length 300. The evaluation parameters for the models were Precision, Recall and F1-score. Micro averages for all evaluation parameters were used due to the balanced class distribution.

Table 1 shows the F1-score of the LSTM, Bi-LSTM and BERT models on embeddings and feature sets. These results show model training after ten epochs. The automated feature extraction using minimum inverse document frequency has shown the highest results for the background rhetoric zone. Background class have the largest number of overlapping vocabulary items with other classes as it contains general context sentences. Similarly, Min_IDF models the most common vocabulary in the feature set. BERT has shown a similar result as Min_IDF for the background zone. Manual features (a total of 2768 features) and Word2vec were not as good as others because the background zone has the highest number of features that were not completely modeled in their case.

BERT model has shown better performance for most classes, i.e., motivation, goal, hypothesis, experiment, and results (as shown in Fig. 5). However, manual features with Bi-LSTM have demonstrated 90% results for the 'problem rhetoric zone'. An analysis of BERT embeddings and manual features for problem class reveals that manual features have bi-gram and tri-gram features such as 'time consuming', 'major barrier', 'remain unstudied' and 'has been neglected' etc. that made classifier more accurate in predicting the class label. Moreover, varying sizes of BERT embeddings need further evaluation, as for the present work, the embedding size was limited to 300 tokens. Whereas the manual feature vector for the problem class contains 292 features. A comparison of the BERT and Word2vec embeddings clearly shows that the underlying SciBERT has given BERT a clear advantage over the Word2vec. Moreover, a t-test evaluation of F1-scores of the BERT (mean 0.633, SD 0.087) compared to Word2vec-LSTM (mean 0.403, SD 0.204) and Word2vec-BiLSTM (mean 0.394, SD 0.202) shows that BERT has performed better i.e. $t(10) = 4.558$, $p = 0.00$ and $t(10) = 4.578$, $p = 0.00$ respectively. Based on the evaluation, the null hypothesis that both models perform the same is rejected in this case as there is a significant difference between both models. Evaluating all models using one-way ANOVA shows a statistically significant difference between models i.e. $F(6,63) = 3.589$, $p = 0.004$. Therefore,

Table 1 F-measure score of the training models (highest value shown as underlined)

Rhetoric zones	Automatic feature extraction					Manual feature extraction	
	Min_IDF		Word2Vec		BERT		
	LSTM	Bi-LSTM	LSTM	Bi-LSTM		LSTM	Bi-LSTM
BACK	<u>0.75</u>	0.70	0.67	0.72	0.73	0.64	0.58
MOTIV	0.34	0.26	0.19	0.18	<u>0.56</u>	0.4	0.42
PROB	0.62	0.61	0.8	0.72	0.79	0.85	<u>0.9</u>
GOAL	0.42	0.36	0.19	0.32	<u>0.55</u>	0.44	0.42
HYP	0.24	0.37	0.5	0.15	<u>0.59</u>	0.57	0.5
MODL	0.41	0.57	0.35	0.29	0.65	<u>0.75</u>	0.70
METH	0.36	0.42	0.32	0.41	0.65	0.66	<u>0.7</u>
EXP	0.46	0.49	0.44	0.48	<u>0.56</u>	0.36	0.47
CON	0.25	0.21	0.23	0.23	0.54	0.53	<u>0.56</u>
RES	0.23	0.33	0.34	0.44	<u>0.71</u>	0.43	0.48

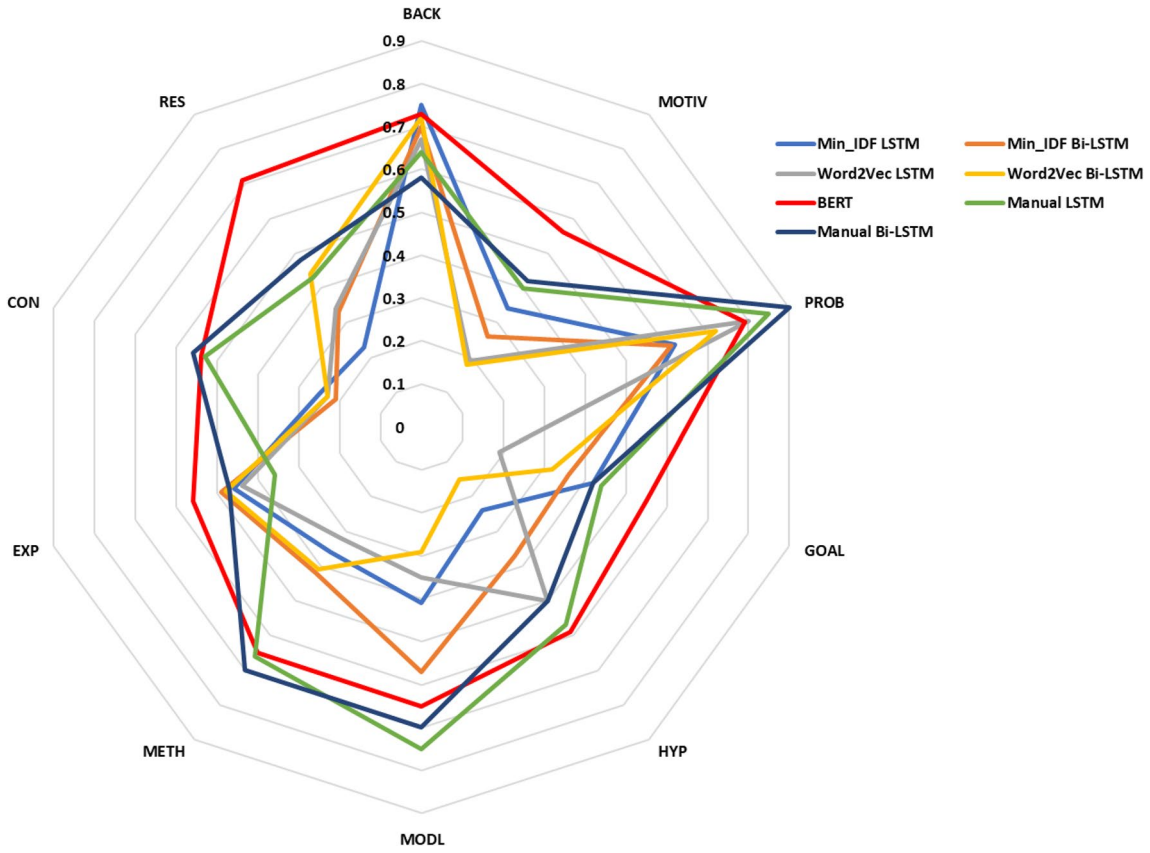

Fig. 5 Comparison of different models (F1-score)

the BERT trained model was selected for classifying the rhetoric zones of research articles during the experiment.

All these models and embeddings were implemented on the Google Colab server with data stored on Google Cloud Storage. Tensorflow deep learning library is used to implement LSTM, Bi-LSTM and BERT models. The web interface for user interaction was implemented using the Flask web framework.

3.4 Rhetoric zones similarity

The present work computes similarity among rhetoric zones after classifying them to generate the final ranked list. Traditional measures for computing similarity are Jaccard, dice, hamming distance and cosine similarity. The problem with traditional approaches is that they compute similarity based on the existence criteria of words that lack contextual similarity. Any negation in a sentence is treated closer to the same as positive. However, recently the similarity of the rhetoric sentences can be computed directly using embeddings, but in this case, the runtime is proportional to the scale of the corpus, i.e., if there are one million sentences or articles in the dataset then one million pairs need to be classified by the deep learning model. To overcome this problem, the present work follows an efficient approach by generating fixed-sized embeddings for every instance of the dataset and embeddings of the incoming query. Both embeddings are then classified according to the rhetoric zones, and finally, the similarity between pairs of classified zones will be computed.

The present work has computed embeddings similarity using a recent unsupervised learning technique called Sent2vec (Pagliardini et al. 2018). Sent2vec is a combination of CBOW model of the Word2vec including n-gram tokens and averaging the embeddings for their summarization to form a single vector in the same latent space. Sent2vec works on distributional hypothesis where words appearing nearby are considered to have the same context. Formally, the Sent2vec learn two embeddings, the source (R_W) and the target (T_W) of h dimension for every word in the vocabulary. Averaging the constituent words of the source word embeddings (R_W) forms the sentence embedding. Sent2vec augmented the source word embeddings by including n-grams (where n = 1,…,n) of each sentence. These n-grams are also averaged along with the words. The Sent2vec embeddings are modelled as formula:

$$E_S = \frac{1}{|NG(S)|} \sum_{w \in NG(S)} R_w$$

The NG(S) is a function that generates the list of n-grams (where n = *1..n*) appears in the sentence (S). Later, the Soft-Max activation function with negative sampling is applied to predict a missing word. Negative sampling is known to be efficient for predicting a large number of output classes. Sent2vec uses binary logistic loss function combined with negative sampling to predict the output class. Sent2vec has a low computational overhead for inference and training as only |S| is required.

4 Experiments and evaluation results

This section presents both the subjective and objective evaluation of the proposed rhetoric zones classification and similarity technique. The subjective evaluation involves domain experts, whilst the objective evaluation is carried out by comparing the proposed technique with the other content-based filtering approaches for citation recommendations.

4.1 Expert-based subjective evaluation

A set of related articles was manually formulated for the experiment. Each of ten senior faculty members from the computer science and biochemistry departments were requested to provide two research articles related to their research domains. These twenty articles are selected as query papers. Against every query paper, each faculty member has provided ten most relevant papers they have reviewed before and are published between 2017 and 2019. Using the keywords from the query papers and also suggested by the faculty members, the top 20 results of different digital libraries (ScienceDirect, PubMed, Wiley, IEEExplore, CORE and Arxiv) were collected. With this method, 2371 articles were gathered that has some relevance to the query papers. Later, the two hundred articles provided by faculty members were added to the experimental set. A very few duplicate articles were removed to form the final set of 2543 articles. Every article was assigned a unique id for identification.

The abstract, introduction section and metadata of articles were extracted manually and stored as text files. Metadata included title, venue and keywords of the article. Sentence-wise tokenization was performed to separate every sentence. Pre-processing such as special characters removal, citation removal and lower-case were applied on the sentence tokens. A total of 118,325 sentences were gathered, with an average of 46 sentences per article.

All these sentences are then classified individually using the proposed model. The model predicts a rhetoric zone for every sentence based on its features. However, sentences with a classification probability of more than 0.5 were considered for similarity comparison. Otherwise, any sentence assigned a rhetoric zone label by the proposed model with a probability less than 0.5 was discarded from further processing. A separate JSON file for every rhetoric zone was created that stored the article id, sentence text and classification probability of the sentence. The same procedure was performed for all twenty query papers, and their JSON files were stored separately. Based on classification and the selection criteria, a total of 16,975 sentences for related papers and 186 sentences for query papers were classified under ten rhetoric zones. The distribution of classified sentences is shown in Table 2.

After classifying the rhetoric zones, the similarities between the individual rhetoric zones of the query papers and the related papers were computed. Based on computed similarity using Sent2vec, the top ten articles were selected in order of descending similarities. The mapping information is marked between the related papers provided by the faculty members and their corresponding query papers, through which the evaluation technique computes the average precision (AP@10) of every query paper for every rhetoric zone. This mapping information is only used for the evaluation of the results retrieved by the proposed algorithm.

Average precision is an evaluation measure that considers both Precision and Recall for ranked retrieval results. Furthermore, average precision gives an indication about the position of the relevant retrieved results in a ranked list by computing the mean of the precision value after each relevant document appears in the list. Usually, average precision is computed for all retrieved documents; however, it can be measured for a given number of results, i.e., known as the cut-off rank or average precision at k denoted by AP@k. Average precision at k (AP@k) considers only the top k results of the ranked list.

$$\text{AP@k} = \frac{1}{gPov} \sum_{i=1}^{k} \text{P@}i \times \text{relevance@}i$$

The equation shows the formula of average precision at k. The k in the equation refers to the number of retrieved documents that shall be considered for evaluation. The gPov is

Table 2 Rhetoric sentence distribution

Articles	N	BACK	MOTIV	PROB	GOAL	HYP	MODL	METH	EXP	CON	RES	Total
Related	2543	3103	2651	1822	1046	776	1127	1084	1989	1512	1865	16,975
Query	20	26	16	11	13	9	14	19	27	23	28	186

the ground truth positive, P@i is the precision at ith item and relevance@i is a function that returns true if the document at the ith position is relevant otherwise false.

Table 3 shows the average precision for the top ten (AP@10) results retrieved by the proposed system. Results were ranked according to rhetoric zone similarity for a given query paper. An evaluation process has evaluated every rhetoric zone of all query papers as shown in Table 3; however, the final ranked list shown to the users is an order according to the average similarity of all rhetoric zones of a query paper. Figure 6 shows the mean average precision (mAP) of the individual rhetoric zone. The highest mAP was achieved by the experiment class whereas, the conclusion class remained at the lowest. It has been observed that

Table 3 Average precision (AP) of query articles @10

Query paper	Rhetoric zones										ALL zones
	BACK	MOTIV	PROB	GOAL	HYP	MODL	METH	EXP	CON	RES	
Q1	0.656	0.656	0.628	0.628	0.511	0.486	0.628	0.433	0.628	0.762	0.602
Q2	0.871	0.678	0.486	0.511	0.628	0.785	0.525	0.385	0.433	0.686	0.599
Q3	0.623	0.475	0.525	0.686	0.489	0.526	0.799	0.628	0.564	0.628	0.594
Q4	0.762	0.762	0.785	0.385	0.628	0.762	0.628	0.686	0.486	0.433	0.632
Q5	0.564	0.511	0.489	0.628	0.785	0.686	0.489	0.762	0.511	0.385	0.581
Q6	0.486	0.486	0.628	0.785	0.799	0.385	0.686	0.564	0.785	0.525	0.613
Q7	0.785	0.852	0.785	0.433	0.489	0.511	0.762	0.511	0.489	0.785	0.640
Q8	0.526	0.526	0.511	0.628	0.433	0.564	0.525	0.525	0.628	0.489	0.536
Q9	0.511	0.413	0.489	0.525	0.385	0.785	0.628	0.785	0.525	0.628	0.567
Q10	0.794	0.799	0.385	0.799	0.564	0.525	0.433	0.886	0.785	0.525	0.650
Q11	0.489	0.475	0.628	0.489	0.628	0.785	0.762	0.762	0.511	0.486	0.602
Q12	0.525	0.525	0.785	0.785	0.564	0.511	0.528	0.785	0.489	0.686	0.618
Q13	0.564	0.486	0.799	0.762	0.686	0.489	0.511	0.628	0.586	0.525	0.604
Q14	0.762	0.686	0.486	0.628	0.785	0.511	0.489	0.489	0.435	0.385	0.566
Q15	0.628	0.628	0.785	0.564	0.385	0.686	0.628	0.799	0.511	0.686	0.630
Q16	0.489	0.433	0.526	0.799	0.489	0.628	0.785	0.785	0.489	0.511	0.593
Q17	0.762	0.762	0.525	0.385	0.486	0.762	0.799	0.511	0.564	0.794	0.635
Q18	0.486	0.528	0.433	0.525	0.785	0.526	0.525	0.489	0.511	0.489	0.530
Q19	0.489	0.511	0.628	0.762	0.526	0.525	0.489	0.433	0.489	0.433	0.529
Q20	0.785	0.525	0.486	0.628	0.564	0.785	0.385	0.762	0.475	0.485	0.588

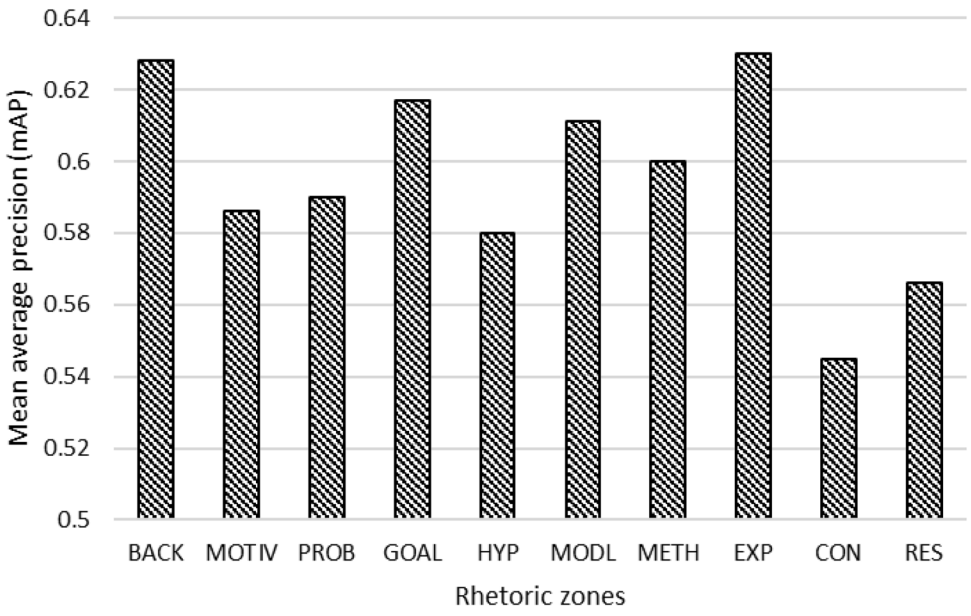

Fig. 6 Mean average precision (mAP) of the rhetoric zones

features of background and conclusion classes overlap with each other, which makes misclassification at several places. Mean average precision provides an indication about a possible weighting scheme that can be assigned to each class for computing similarity.

Average precision is an objective evaluation method in which an automated process measures the performance of the recommendations made by the proposed system. In addition to objective evaluation, the normalized discounted cumulative gain (nDCG), a subjective evaluation measure was computed involving ten experts. NDGC evaluates the system based on the graded relevance of the retrieved results. Grading is performed by the experts on a Likert scale. In the present case, all ten faculty members were experts and the Likert scale range between 1 and 3, with 1 as highly relevant and 3 for the low. Every faculty member is provided with a list of the top ten results retrieved by the proposed system against their provided query paper. A web-based interface is provided to the faculty members for browsing the results and viewing the complete article if required. The equation below shows the formula of nDCG.

$$nDCG_K = \frac{1}{IDCG_K} \times \sum_{i=1}^{K} \frac{2^{rel_i} - 1}{log_2(i+1)}$$

The nDCG formula is the product of normalizing factor on the left and the discounted cumulative gain (DCG) on the right. The k represents the rank position to limit the varying length of the results. The DCG penalized the score if the highly relevant document appears at the lower rank in the result list. The rel*i* is the grading value assigned by an expert to the document at ith position. The value of nDCG (normalized discounted cumulative gain) ranges between 0 and 1, with 1 as the ideal ranking. Ideal discounted cumulative gain ($IDCG_K$) is the maximum possible discounted cumulative gain (DCG) at kth position. Table 4 shows the results of nDCG@10 of the proposed rhetoric zone classification and similarity technique.

The nDCG results are better than AP@10 for respective query papers. This shows that the articles retrieved by the proposed technique are relevant papers. From the average precision and nDCG results, it can be concluded that proposed system has retrieved related articles among the top 10 results that are either provided by the faculty member or available in the dataset. The overall impression of the faculty members who used and evaluated the proposed system was "efficient" as most of them have commented that few articles the proposed system has retrieved are highly relevant to the query paper and they themselves have not found those articles before.

4.2 Comparison with other approaches—objective evaluation

The performance of the proposed rhetoric zone classification and similarity model (RtZone) has been compared with several other citation recommendations models that are based on the content filtering approach. These models are summarized as follows:

NNRank (Bhagavatula et al. 2018): Neural network ranking is content-based filtering method that represents query and candidate documents into vector space model and use the nearest neighbour technique to rank the relevant results. For the present experiment, the hyperparameters such as size of embeddings (dim = 325) and hidden layers ($dim_H = 150$), regularization strength ($\lambda_1 = 0$, $\lambda_2 =$ 1e-6), number of epochs (epo = 256) and learning rate (0.001) are kept the same as the original study.

LSTM-CAV (Wang et al. 2020a): The LSTM based personalized context-aware citation recommendation model is a hybrid approach using both collaborative and content-based filtering. LSTM-CAV process author, venue, and keywords information along with the content of the article in the form of distributed vector representation. The LSTM model learns both the article and the citation contexts and then measure relevance between them. A ranked list is returned based on high relevance scores. LSTM-CAV was evaluated on the AAN dataset using an embedding size of 150, regularization weights for $\lambda_1 =$ 1e−5 and $\lambda_2 =$ 1e−6 with stochastic optimization method AdaGrad having a learning rate set to 0.001.

Doc2vec model (Le and Mikolov 2014): The D2V model is an unsupervised learning technique for representing documents as fixed-length dense feature vector representation. D2V, as compared to the traditional bag-of-words approach, learns the feature vector using a neural network and keep in view the ordering among words and individual word semantics. Similarity among documents is measured by their representational relevance score. For the present evaluation, the embedding size is kept at 150 as consistent to Doc2vec original experiment.

Table 4 Results of normalized cumulative discounted gain (nDCG@10)

Query paper	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
nDCG	0.728	0.740	0.743	0.755	0.718	0.782	0.699	0.724	0.801	0.722
Query paper	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20
nDCG	0.626	0.709	0.664	0.672	0.587	0.665	0.718	0.721	0.634	0.680

GRU-MTL (Bansal et al. 2016): A latent vector of text sequences is encoded using gated recurrent neural units (GRUs) for citation recommendation on the collaborative filtering task. Full text of articles is used here for evaluation with embedding size 200 and dimensions of hidden layers of first and second recurrent neural network (RNNs) as $dim_{H1} = 200$ and $dim_{H2} = 400$, respectively.

Scholarfy (Achakulvisut et al. 2016): The vectorization of a document is carried out by Latent Semantic Analysis (LSA) with a combination of log-entropy and Tf-idf for weighting purpose. Scholarfy has used abstracts of the articles. Recommendations are made using Rocchio Algorithm for finding nearest neighbour articles. An embedding or vector size of 150 is selected here for evaluation that is the same as the original experiment carried out by the Scholarfy.

The real-world bibliographic dataset named the ACL anthology network (AAN) (Radev et al. 2013) is used for the performance evaluation of the proposed model as compared to previous content-based citation recommendation approaches. AAN contains articles on natural language processing (NLP) and computational linguistics collected from different venues. After removing papers with incomplete information, the dataset for evaluation contains 27,324 papers. Abstract and introduction section is used for the evaluation of proposed model whereas abstract only for Scholarfy and full text for all other techniques. LSTM-CAV is provided with author, venue and keyword information in addition to the full text of the articles.

The comparison of the proposed rhetoric zones similarity model against the baseline approaches in terms of Recall is shown in Fig. 7. The comparison results show that the recommendations made by the proposed model are more precise as compared to all other baseline approaches. The reason that the proposed model computes similarity among articles is based on the semantics of their content rather than considering individual words appearing in the text as Doc2vec, Scholarfy and NNRank. The NNRank has shown second best results for AAN dataset; however, it has not shown the similar in case of ART + CORE dataset because it is well known that the nearest neighbour approach is highly sensitive to irrelevant features and scale of data. The AAN dataset mainly contains papers on a specific topic, whereas the ART + CORE has general computer science articles. Moreover, the top 325 features learned by the NNRank in the case of AAN are all relevant to candidate articles; however, it does not remain the same for multi-domain articles. Similar behavior as NNRank is shown by Scholarfy with its LSA approach. It remains consistent with below average results for both datasets due to its dependency on representations provided by the LSA. Comparatively, the LSTM-CAV and GRU-ML generate insignificant results due to their small embedding size against full article text. LSTM-CAV have reported the same, as a limitation of their work that increasing the embedding size to 500 and 1000 in some cases produce better results. A conclusion can be made from these results that the proposed method based on semantic similarity has outperformed the state-of-the-art content-based filtering methods using syntactical similarity for citation recommendations.

5 Conclusion

In the present work, a citation recommendation system using deep learning technique is proposed, which considers both local and global context-aware citation recommendations approach and presents a remedy to the cold-start problem. Previous research has heavily addressed the cold-start problem using collaborative filtering techniques relying on pre-computed or available information about articles. However, the present proposal is based on content filtering, which requires no prior information about query

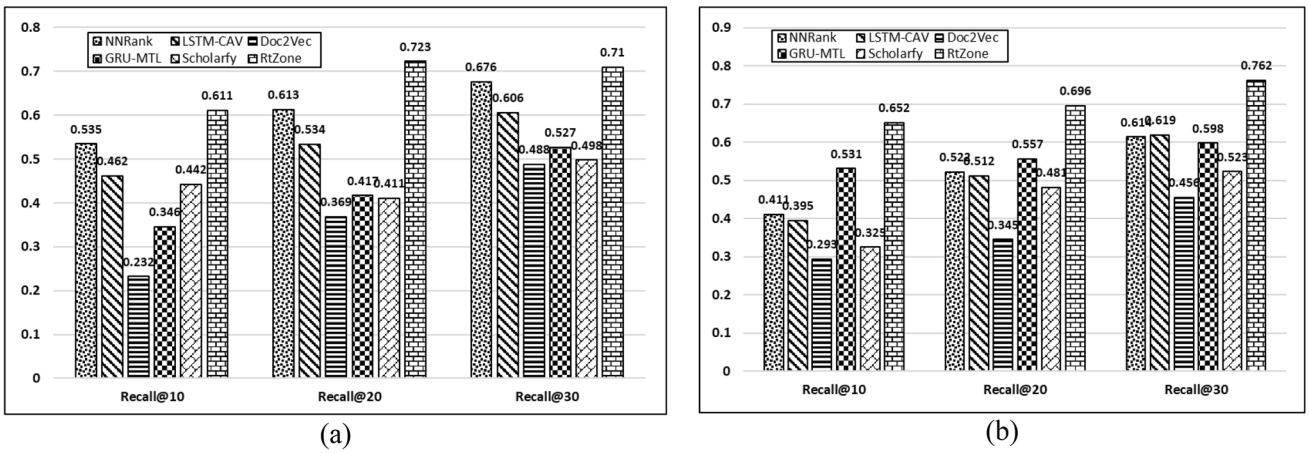

Fig. 7 Recall on the **a** AAN dataset **b** ART + CORE dataset

and candidate articles. Citation recommendations are made through rhetoric zones classification using Bi-LSTM and BERT models and computing similarity using Sent2vec embeddings of every individual zone. Moreover, Metadata information is combined with rhetoric zone information to produce better results. The proposal is both an offline and online approach that computes article relatedness based on the semantics of the content as a solution to cold-start problem. The deep learning model was trained using well-known ART and CORE datasets. The trained model with an accuracy of over 80% is tested on 2,543 articles. The objective evaluation using mean average precision and subjective evaluation using normalized discounted cumulative gain with ten experts was computed. The evaluation results clearly show the effectiveness of the proposed approach in terms of citation recommendation.

The proposed approach is validated through experiments for citation recommendation. However, there are several limitations that need further study. Currently, all rhetoric zones have been assigned the same weight during similarity computation. During this research, it has been observed that setting dynamic weights to rhetoric zones shall improve the final ranking of the recommendation list. Moreover, the size of embeddings window is kept to 300, which can be increased or decreased to further evaluate its effectiveness. Sometimes rhetoric zones overlap each other; currently the present work assigns a single class label to a zone, whereas multi-class classification might result differently. The present word considers the rhetoric zone of sentence length from the introduction section only, which can be extended to multiple sentences and multiple sections of the paper. The current approach carried an experiment with a collection of articles to show its validity. However, it can be compared to top-ranking article recommendations of search systems such as Google Scholar, ScienceDirect, CORE and DBLP.

Data availability The data can be obtained by sending request e-mails to the corresponding author.

Declarations

Conflict of interest There is no potential conflict of interest in this study.

References

Abro WA, Qi G, Ali Z, Feng Y, Aamir M (2020) Multi-turn intent determination and slot filling with neural networks and regular expressions. Knowl-Based Syst 208:106428. https://doi.org/10.1016/j.knosys.2020.106428

Achakulvisut T, Acuna DE, Ruangrong T, Kording K (2016) Science concierge: a fast content-based recommendation system for scientific publications. PLoS ONE 11(7):1–11. https://doi.org/10.1371/journal.pone.0158423

Ali Z, Kefalas P, Muhammad K, Ali B, Imran M (2020a) Deep learning in citation recommendation models survey. Expert Syst Appl 162:113790. https://doi.org/10.1016/j.eswa.2020.113790

Ali Z, Qi G, Muhammad K, Ali B, Abro WA (2020b) Paper recommendation based on heterogeneous network embedding. Knowl-Based Syst 210:106438. https://doi.org/10.1016/j.knosys.2020.106438

Ambalavanan AK, Devarakonda MV (2020) Using the contextual language model BERT for multi-criteria classification of scientific articles. J Biomed Inform 112:103578. https://doi.org/10.1016/j.jbi.2020.103578

Asadi N, Badie K, Mahmoudi MT (2019) Automatic zone identification in scientific papers via fusion techniques. Scientometrics 119(2):845–862. https://doi.org/10.1007/s11192-019-03060-9

Badie K, Asadi N, Tayefeh Mahmoudi M (2018) Zone identification based on features with high semantic richness and combining results of separate classifiers. J Inform Telecommun 2(4):411–427

Bai X, Wang M, Lee I, Yang Z, Kong X, Xia F (2019) Scientific paper recommendation: a survey. IEEE Access 7:9324–9339. https://doi.org/10.1109/ACCESS.2018.2890388

Bansal T, Belanger D, McCallum A (2016) Ask the GRU: multi-task learning for deep text recommendations. In: Proceedings of the 10th ACM Conference on Recommender Systems. Association for Computing Machinery, New York, NY, USA, pp. 107–114. https://doi.org/10.1145/2959100.2959180

Bär D, Zesch T, Gurevych I (2011) A reflective view on text similarity. In: Proceedings of the International Conference Recent Advances in Natural Language Processing 2011 pp 515–520

Bhagavatula C, Feldman S, Power R, Ammar W (2018) Content-based citation recommendation. In: Proceedings of the 2018 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Association for Computational Linguistics, New Orleans, Louisiana, pp 238–251. https://doi.org/10.18653/v1/N18-1022

Blagus R, Lusa L (2013) SMOTE for high-dimensional class-imbalanced data. BMC Bioinform 14(1):106. https://doi.org/10.1186/1471-2105-14-106

Christoforidis G, Kefalas P, Papadopoulos A, Manolopoulos Y (2018) Recommendation of points-of-interest using graph embeddings. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), pp 31–40. https://doi.org/10.1109/DSAA.2018.00013

Chughtai GR, Lee J, Shahzadi M, Kabir A, Hassan MAS (2020) An efficient ontology-based topic-specific article recommendation model for best-fit reviewers. Scientometrics 122(1):249–265. https://doi.org/10.1007/s11192-019-03261-2

Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding

Habib R, Afzal MT (2019) Sections-based bibliographic coupling for research paper recommendation. Scientometrics 119(2):643–656. https://doi.org/10.1007/s11192-019-03053-8

Han J, Song Y, Zhao WX, Shi S, Zhang H (2018) hyperdoc2vec: distributed representations of hypertext documents. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (volume 1: long papers). Association for Computational Linguistics, Melbourne, Australia, pp 2384–2394. https://doi.org/10.18653/v1/P18-1222

Jang B, Kim M, Harerimana G, Kang S, Kim JW (2020) Bi-LSTM model to increase accuracy in text classification: combining Word2vec CNN and attention mechanism. Appl Sci 10(17):5841. https://doi.org/10.3390/app10175841

Jeong C, Jang S, Park E, Choi S (2020) A context-aware citation recommendation model with BERT and graph convolutional networks. Scientometrics 124(3):1907–1922. https://doi.org/10.1007/s11192-020-03561-y

Knoth P, Anastasiou L, Charalampous A, Cancellieri M, Pearce S, Pontika N, Bayer V (2017) Towards effective research recommender systems for repositories. In: Open Repositories 2017

Kumar V, Recupero DR, Riboni D, Helaoui R (2021) Ensembling classical machine learning and deep learning approaches for morbidity identification from clinical notes. IEEE Access 9:7107–7126. https://doi.org/10.1109/ACCESS.2020.3043221

Le QV, Mikolov T (2014) Distributed representations of sentences and documents. Distributed representations of sentences and documents, pp 1188–1196

Li X, Chen Y, Pettit B, Rijke MD (2019) Personalised reranking of paper recommendations using paper content and user behavior. ACM Trans Inf Syst. https://doi.org/10.1145/3312528

Liakata M, Soldatova L (2009a) The ART corpus. Retrieved from https://www.aber.ac.uk/en/cs/research/cb/projects/art/art-corpus/

Liakata M, Soldatova LN et al. (2009) Semantic annotation of papers: interface & enrichment tool (sapient). In: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, pp 193–200

Ma X, Wang R (2019) Personalized scientific paper recommendation based on heterogeneous graph representation. IEEE Access 7:79887–79894. https://doi.org/10.1109/ACCESS.2019.2923293

Ma X, Zhang Y, Zeng J (2019) Newly published scientific papers recommendation in heterogeneous information networks. Mobile Netw Appl 24(1):69–79. https://doi.org/10.1007/s11036-018-1133-9

Ma S, Zhang C, Liu X (2020) A review of citation recommendation: from textual content to enriched context. Scientometrics 122(3):1445–1472. https://doi.org/10.1007/s11192-019-03336-0

Mahdi MN, Ahmad AR, Ismail R, Natiq H, Mohammed MA (2020) Solution for information overload using faceted search—a review. IEEE Access 8:119554–119585. https://doi.org/10.1109/ACCESS.2020.3005536

Martins GB, Papa JP, Adeli H (2020) Deep learning techniques for recommender systems based on collaborative filtering. Expert Syst 37(6):e12647. https://doi.org/10.1111/exsy.12647

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space

Pagliardini M, Gupta P, Jaggi M (2018) Unsupervised learning of sentence embeddings using compositional n-gram features. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). https://doi.org/10.18653/v1/n18-1049

Radev DR, Muthukrishnan P, Qazvinian V, Abu-Jbara A (2013) The ACL anthology network corpus. Lang Resour Eval 47(4):919–944. https://doi.org/10.1007/s10579-012-9211-2

Si Y, Wang J, Xu H, Roberts K (2019) Enhancing clinical concept extraction with contextual embeddings. J Am Med Inform Assoc 26(11):1297–1304. https://doi.org/10.1093/jamia/ocz096

Wang J, Zhu L, Dai T, Wang Y (2020a) Deep memory network with Bi-LSTM for personalized context-aware citation recommendation. Neurocomputing 410:103–113. https://doi.org/10.1016/j.neucom.2020.05.047

Wang W, Liu J, Tang T, Tuarob S, Xia F, Gong Z, King I (2020b) Attributed collaboration network embedding for academic relationship mining. ACM Trans Web. https://doi.org/10.1145/3409736

Zeng T, Acuna DE (2020) Modeling citation worthiness by using attention-based bidirectional long short-term memory networks and interpretable models. Scientometrics 124(1):399–428. https://doi.org/10.1007/s11192-020-03421-9