Hindawi
Complexity
Volume 2021, Article ID 2893257, 8 pages
https://doi.org/10.1155/2021/2893257

WILEY | Hindawi

Research Article

Impact of Lexical Features on Answer Detection Model in Discussion Forums

Atif Khan,[1] Muhammad Adnan Gul,[1] Abdullah Alharbi,[2] M. Irfan Uddin,[3] Shaukat Ali,[1] and Bader Alouffi[4]

[1]*Department of Computer Science, Islamia College Peshawar, Peshawar, KP, Pakistan*
[2]*Department of Information Technology, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia*
[3]*Institute of Computing, Kohat University of Science and Technology, Kohat, Pakistan*
[4]*Department of Computer Science, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia*

Correspondence should be addressed to Atif Khan; atif.softeng@gmail.com

Received 13 July 2020; Accepted 31 March 2021; Published 15 April 2021

Academic Editor: Ning Cai

Online forums have become the main source of knowledge over the Internet as data are constantly flooded into them. In most cases, a question in a web forum receives several responses, making it impossible for the question poster to obtain the most suitable answer. Thus, an important problem is how to automatically extract the most appropriate and high-quality answers in a thread. Prior studies have used different combinations of both lexical and nonlexical features to retrieve the most relevant answers from discussion forums, and hence, there is no standard/general set of features that could be effectively used for relevant answer/reply post classification. However, this study proposed an answer detection model that is exclusively relying on lexical features and employs a random forest classifier for classification of answers in discussion boards. Experimental results showed that the proposed answer detection model outperformed the baseline technique and other state-of-the-art machine learning algorithms in terms of classification accuracy on benchmark forum datasets.

1. Introduction

Web forum is a virtual online network of like-minded individuals where they collaborate with each other. The collaboration starts when a user asks question and others answer it. Usually, a question receives several answers and that makes it difficult to extract an appropriate answer to the question for the question poster. Thus, an important problem is how to automatically extract the most appropriate and high-quality answers in a thread, and it needs to be resolved in order to avoid the laborious and tedious task of scanning all the replies manually.

Basically, the extraction of best answers/replies is a classification task [1–5]. Replies are distributed into non-quality class, low-quality class, and high-quality class based on its importance to the question being asked. In order to classify the reply, it is essential to judge the reply content quality. From reply content quality, we mean to what extent/degree an answer responds to the query or question. Usually, different types of features, explained below, are used to assess the quality of reply content.

1.1. Syntactic. Features extracted from sentence elements and their structure are called syntactic features [6]. Mainly three approaches: parts of speech (POS), order of words, and sentence grammar, are used to extract these features.

1.2. Lexical. These are string-based similarity features in which sentences are considered as character sequences [7];

for example, cosine similarity of two documents is calculated between their feature vectors.

1.3. Semantic. These features are used for contextual similarity of text documents. These are helpful when sentences have different words with similar meanings. Semantic similarity sees sentences in their context. Different approaches are used to find semantic similarity. Corpus-based approach is based on statistical analysis; knowledge-based approach utilizes an external resource called WordNet to measure the relatedness/similarity of two words. In WordNet, English words are grouped into synsets, which are organized into a hierarchy forming a semantic network in which semantic relations between synsets can be obtained easily. Each synset in the semantic network represents a group of synonyms and has a single meaning. The third approach is structure-based, which exploits the sentence structure to measure the sentence similarity; that is, similar sentences have a similar basic structure [7].

In the context of discussion forums, there are some additional features related to forums that are author-related, answer response time, document contents, and thread structure.

1.4. Forum-Related. These are forum-specific features also called forum metadata, for example, rating, liking, disliking, or voting mechanism of a reply/answer post.

1.5. Thread-Related. It comprises features that are related to the thread structure. For instance, who is the question poster? Have a reply mentioned another user/author. Has the reply been given by the question poster? Is a reply referring to another reply?

1.6. Timeliness. It is time elapsed between the question and a given reply or time elapsed between two replies, etc.

1.7. Author-Related. It represents the reputation and activeness of an author in the forum. Different rating mechanisms are used in forums to assign some specific values to authors. This shows their expertise level and authenticity.

1.8. Content-Based. These features represent the contents of a sentence. For example, does a text have some special strings like hyperlinks and 5W1H words?

We categorize the above features into two types, lexical and nonlexical, as shown in Table 1. Lexical features are further classified into syntactic, string-based, and semantic features that are used for general text classification [7] and specifically for answer relevancy/similarity with the given question in discussion forums [8–12]. The authors in [8] used both lexical and nonlexical types to classify reply post as non-quality, low quality, and high quality. The authors in [13] preferred only nonlexical features for answer extraction, while some studies used only lexical features for the same task. Thus, prior studies have used different combinations of both nonlexical and lexical features, and hence, there is no standard/general set of features that could be effectively used for text classification.

However, this study proposes an answer detection model that is totally relying on lexical features and employs the random forest classifier for answer retrieval in discussion forums. In this work, we thoroughly examine the impact of lexical features on relevant answer retrieval in discussion forums in the context of the proposed answer detection model and other advanced classification models.

Since we examine the effect of lexical features on the answer detection model, the feature set of the proposed model includes only lexical features and no nonlexical features (forum, author, thread structure, or time-related) are included; thus, the model is generic and can be used for answer extraction in any type of discussion forums and text classification/relevancy-based problem. We test the model with two datasets: Ubuntu which is a technical forum and general discussion forum TripAdvisor (NYC).

The proposed answer detection model has many potential applications. Since it does not include any forum-specific or data-dependent features and exclusively relies on lexical features, which are set of generic/independent features, the model can be used for answer retrieval in any text classification/relevancy and discussion forum-based problem. Question/answer forums like Yahoo! and Answer 1 could use the proposed model to suggest answers to their users by retrieving them from forum threads. With the proposed model, we can also produce question-answer pairs, which can be further narrowed down to frequently asked questions (FAQs). FAQs can be further used to enrich chatbot knowledge. Contribution of the proposed work is given as follows:

(a) To propose an answer detection model based on generic lexical features and a random forest classifier in order to examine the impact of lexical features for answer retrieval in discussion forums.

(b) To evaluate the effectiveness of the proposed answer detection model in the context of TripAdvisor (NYC) and Ubuntu datasets.

The rest of the paper is organized as follows: Section 2 is about related work. Proposed methodology is explained in Section 3. Section 4 presents experimental settings, results, and discussion. Section 5 illustrates conclusion as well as future work.

2. Related Work

The extraction of most relevant and quality replies/answers is a text classification problem [1–5]. Reply posts are classified as non-quality, low quality, and high quality, on the basis of its relevancy with the question. To classify the text, it is necessary to judge its quality [14]. For quality judgment, different types of features are used. Since all features are not equally important, features that are nonvaluable and redundant are eliminated [8, 15] using different features selection/reduction techniques such as chi-squared (CHI),

Table 1: Lexical and nonlexical features.

Type	Features	Description
Lexical	Syntactic	Extracted from sentence elements and structure
	String-based	String/words base focusing on words of sentences
	Semantic	Different words with the same meaning (contextual similarity)
Nonlexical	Forum-related	Forum-specific features
	Thread-related	Features related to thread structure
	Timeliness	Time elapsed between question and reply post
	Author activeness	Shows how an author is active in the forum
	Content-based	Does the text have some specific words/strings?

information gain (IG), document frequency thresholding (DF), Acc and Acc2 metrics, and univariate and clustering features [16, 17].

The study proposed in [8] is closely related to our work. They classified reply posts in online discussion forums into non-quality, low quality, and high quality by taking into account the reply-post relevancy with the question. Moreover, they grouped features used for reply-posts classification into six classes: ease of understanding, author activeness, amount of data, politeness, relevancy, and timeliness. They were separated further into twenty-eight (28) lexical and nonlexical features. They used various selection techniques to minimize feature space in order enhance the model performance. The authors in [13] studied five feature groups that are content-based, lexical, structural, reply-to, and forum-specific, in order to determine the candidate answers quality to a question post in discussion forums. The five groups of features were further divided into subfeatures. Some researchers [13] suggested that the lexical similarity of the answers with the question is minimum, and in such case, nonlexical features are vital and more reliable to judge the quality of contents [18].

The authors in [7] used nonlexical with n-grams of lexical features, while the authors in [1] employed nonlexical features (user interactive behaviour) for the classification of massive open online course (MOOC) threads using the deep learning technique. Since it is a user interactive behaviour-based model, this makes it content and language independent. The authors in [19] used content-based and structural features for < title and reply > pairs retrieval to improve the chatbot knowledge. The authors in [12] classify online forum threads by utilizing nonlexical features into subjective and nonsubjective. The authors in [9] proposed a model for patterns extraction from questions and nonquestions and uses the extracted patterns for classification of forum text into question and nonquestion. Finally, a graph-based approach was used for the retrieval of answers in the same thread.

Bag-of-Word technique with cooccurrences features (contextual similarity features) from Wikipedia [20] is also exploited to classify news articles into twenty groups. A study conducted in [21] used lexical and nonlexical features (question words, forum metadata, and a basic question mark rule) for extraction of question in web forums. A researcher in [12] classified online forum threads into subjective and nonsubjective classes using nonlexical thread-specific features.

Lexical and semantic features were used in [22] to classify short text, while the authors in [6] used semantic features to find similarity between academic articles. The authors in [23] proposed a model for paraphrase identification in news articles in order to avoid news about similar events using lexical, syntactic, and semantic features. Some studies [24, 25] used WordNet-based, Word2Vec-based, corpus-based, alignment-based, and literal-based features to find semantic similarity of short English sentences.

Various machine learning algorithms such as support vector machine (SVM), Naïve Bayes (NB), multimodal deep nets, and convolutional neural networks (CNNs) have been exploited for retrieving quality information from the online forums [1, 4, 8, 17, 26–28].

From the above literature, it is clear that there is no standard set of features that can be used for text classification specifically for answer retrieval in online web forums. Lexical, nonlexical, and its different combinations have been utilized to extract quality information (contents) from the discussion forums. Different combinations of features are attempted depending on the nature of forum data. Moreover, mostly forum/data-dependent (nonlexical) features have been used, which make the models forum or data dependent.

In order to address the underlying issues, we propose a forum-independent answer detection model that utilizes lexical features and the random forest classifier for detection of answer in web forums. Lexical features are the set of forum/data-independent features that can be used for answer extraction in any type of discussion forum. The lexical features are string-based, semantic, and syntactic features, and they can also be used for any text classification problem. For fair evaluation, the proposed model is evaluated in the context of both technical discussion forum (Ubuntu) and general discussion forum TripAdvisor (NYC). In addition, the performance of the proposed model is compared with state-of-the-art classifiers for retrieval of answer in the online forums. The next section demonstrates the methodology of the proposed model in detail.

3. Proposed Methodology

The framework of the proposed answer detection model for relevant answer extraction in discussion forums is depicted in Figure 1. It consists of three steps. In the first step, forum data are cleaned by applying preprocessing techniques.

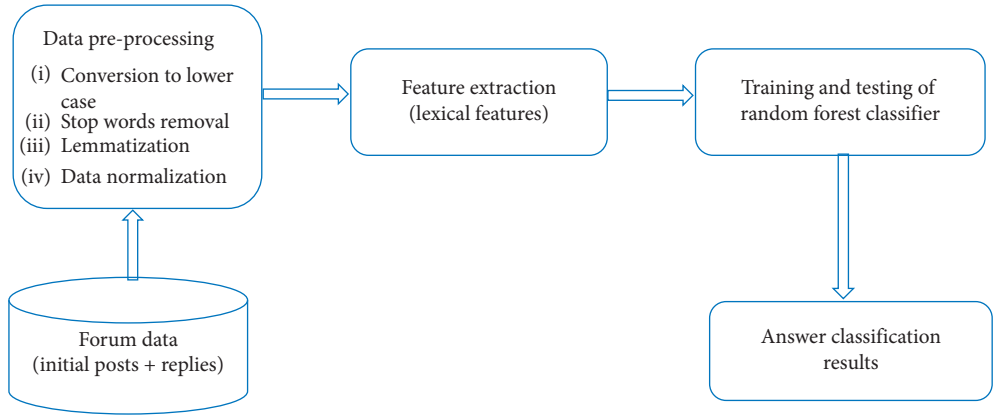

Figure 1: Proposed lexical feature-based answer detection model for answer extraction in discussion forums.

Then, lexical features are extracted from cleaned forum data, and hence, forum data are represented by feature vectors. Finally, the proposed model employing the random forest classifier is trained and tested with forum data, and the results are compared with the state-of-art classifiers.

3.1. Preprocessing. Preprocessing is the process of transforming raw data into an analyzable and predictable format. Steps taken for data preprocessing are as follows:

The forum data are split into tokens (words) using Whitespace Tokenizer (), and then, words are lemmatized to their root form using WordNet Lemmatizer. We used Natural Language Toolkit (NLTK) library for tokenization and lemmatization.

All the words in the forum data are then converted to lowercase.

Next, stop words are removed from the forum data using a list of stop words from NLTK library. They are words that carry little meaning in the sentence such as "a," "an," "the," and "them."

Finally, the forum data are normalized data using Min-Max Scaler of Sklearn library.

3.2. Feature Extraction. The goal of this step is to extract lexical features given in Table 2 from the forum data and hence represent the forum data by feature vectors. The features F1, F2, and F3 in Table 2 are the textual similarities of a reply with title, initial post, and thread centroid, respectively; and in order to calculate these features, we need a numeric representation of forum text data. In this work, we employ Bag-of-Word (BoW) approach to create feature vectors from forum text data and cosine similarity from Sklearn library for calculating similarities between feature vectors. Bag-of-Word (BoW) is a well-known technique/approach [29] used for vector representation of text documents. In this technique, all unique words/features are extracted, and values are assigned based on their frequencies in the document.

In the default behaviour of BoW approach, single words are taken as features, which are called unigrams. Here, feature order and sentence structure are ignored, and only feature frequency is taken into consideration. To overcome this deficiency, a higher order of word sequences, bigrams, and trigrams are used, in which more than one word are taken as features. Bigrams and trigrams give more meaning and information from the document.

Some features are extremely frequent, but they are less important and are nonvaluable. To filter out these features, the term frequency-inverse document (TF-IDF) scheme is used. TF-IDF assigns values/weightage to features based on their importance in the forum document.

Features F4, F5, F6, and F7 given in Table 2 are the features that are calculated directly from the text. Feature F4 represents common words of thread title and reply post. First, title and reply text are split into a sequence of words, and then, common words are extracted. Similarly, feature F5 shows common words of question and reply post and is calculated in a similar way. Feature F6 is the number of reply-post words, which is extracted through simple len function in *Python* language, while F7 contains total unique words in a reply, calculated by using len and set functions in *Python* language.

Features F8, F9, and F10 are the contextual similarities of thread title and reply, question and reply post, and thread centroid and reply post, respectively. A pretrained Google Word2Vec model and word mover distance (WMD) are used for determining contextual similarities. Google's pre-trained Word2Vec model is used for determining semantic/contextual similarity between words, and it has over three million of words/phrase vectors and is trained on some 100 billion words from Google News. WMD measures the contextual dissimilarity of two text documents. Dissimilarity is directly proportional to WMD. Two documents are said to be completely relevant to each other if the WMD value is zero.

3.3. Classification Algorithms. We chose a random forest classifier to a new problem that is reply-post classification in web forum threads. It is extensively used to address the

Table 2: A brief description for lexical features.

Code	Abbreviation	Description
F1	title_reply_cosin_similarity	Cosine similarity between reply post and thread title
F2	question_reply_cosin_similarity	Cosine similarity of an initial post with the reply post
F3	thread_centroid_reply_cosin_similarity	Cosine similarity between thread centroid and reply
F4	Reply_words_overlapping_thrd_title	Thread title and reply common words ratio with the reply words
F5	reply_words_overlapping_initial_post	Question and reply common words ratio with the reply words
F6	total_num_of_words_in_reply	Total number of reply words
F7	unique_words_in_reply	Number of unique reply words
F8	title_reply_wmd	Thread title and reply word mover distance
F9	question_reply_wmd	Question and reply word mover distance
F10	thread_centroid_reply_wmdistance	Word mover distance between thread centroid and reply post

problem of text classification [27]. The performance of random forest classifier is also compared with other baseline and state-of-art text classification algorithms. The classification models are concisely described as follows.

3.3.1. Random Forests (RFs). These classification models are also called random decision forests. For the purpose of classification task, they used an ensemble learning method by constructing number of decision trees during training time and only give those classes as an output that are mode of the classes predicted by the individual decision trees.

3.3.2. Naïve Bayes (NB). These classification models are supervised learning algorithms based on Bayes theorem. It assumes that every feature is independent of every other feature. These classification models are widely used in the area of text classification and show promising results [28].

Bayes theorem is stated as follows:

$$P(y|x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots, x_n|y)}{P(x_1, \ldots, x_n)}, \quad (1)$$

where y is the class variable and x_1 to x_n are dependent features vector.

A small amount of data are required for training this model. Naïve Bayes classifier is very fast, when compared to other classification models. The proposed approach employed multinomial Naïve Bayes which is a variant of NB classifier.

3.3.3. Support Vector Machine (SVM). This group of supervised learning algorithms is used for regression, outlier detection, and classification-related task. It uses less memory and performs efficiently in high-dimensional space [11]. It uses different kernels, but a custom kernel can also be specified. Following two variants were used in this study.

3.3.4. Support Vector Classification (SVC). It is a *libsvm*-based classification model. The fit time of the SVC rises quadratically as the number of samples increases. The default kernel of SVC is "rbf." Other kernels are "sigmoid," "poly," and "linear."

3.3.5. Linear SVC. It is a "liblinear"-based classification model that uses "linear" kernel.

An input may be dense or sparse and is more flexible in selecting penalty or loss functions.

3.3.6. Logistic Regression (LR). It is a classification model that uses generalised logistic regression (LR) to solve multiclass problems with more than two discrete outcomes. It uses a set of input features to predict the probabilities of different outcomes of a target variable.

4. Experimental Settings

4.1. Evaluation Data. The proposed model is evaluated using two datasets—technical one, Ubuntu Linux distribution forum (http://ubuntuforums.org), and nontechnical, online TripAdvisor forum (https://www.tripadvisor.com.my/ShowForum-g28953-i4-New_York.html) for New York City (NYC). Hundred discussion threads were randomly selected from both datasets. Each thread consists of an initial post (question) and replies (answers). Replies having high relevancy with initial post were classified as high quality with class label 3, partially relevant were categorized as low quality with class label 2, and irrelevant replies were classified as nonquality with label 1. The thread structure consists of seven columns: "ThreadID," "Title," "UserID_inipst," "Questions," "UserID," "Replies," and "Class." There are a total of 756 replies in the Ubuntu dataset and 788 replies in the TripAdvisor (NYC) dataset. 80% of data were used for training and 20% for testing purposes.

4.2. Experimental Results and Discussion. The performance of the proposed answer detection model using random forest classifier is compared with 4 state-of-the-art classifiers that are LinearSVC, SVC, logistic regression, and MultinomialNB. These classifiers were tested with lexical features extracted from both datasets (Ubuntu and TripAdvisor-NYC).

In the first phase, all the lexical features given in Table 2 are extracted from the Ubuntu dataset, and hence, data are represented by feature vector representation. Then, the proposed model and other state-of-art classifiers are trained and tested on given data using 10-fold cross-validation. The percentage classification accuracy of all classifiers on lexical

Table 3: Classification accuracy on lexical features for the Ubuntu dataset.

Classifier	Accuracy (%)
Random forest	95.4
SVC	73.0
Logistic regression	67.7
LinearSVC	67.1
MultinomialNB	66.4

Table 4: Classification accuracy on lexical features for the TripAdvisor (NYC) dataset.

Classifier	Accuracy (%)
Random forest	95.6
SVC	73.4
Logistic regression	62.7
LinearSVC	61.4
MultinomialNB	57

features extracted from the Ubuntu dataset is shown in Table 3.

It can be observed from the results given in Table 3 that all classifiers performed well, but random forest performed effectively well and gave the highest accuracy of 95.4%. SVC also performed well and resulted in 73% accuracy. Logistic regression gave 67.7% accuracy. LinearSVC accuracy was 67.1, while MultinomialNB gave the lowest accuracy of 66.4%.

In the next phase, all the lexical features given in Table 2 are extracted from the TripAdvisor (NYC) dataset, and hence, the NYC dataset is represented by feature vector representation. Then, the proposed model and other state-of-the-art classifiers are trained and tested on the NYC dataset using 10-fold cross validation. The results of classification accuracy of all classifiers for lexical features extracted from NYC datasets are shown in Table 4.

It can be seen from the results in Table 4 that the proposed answer detection model using the random forest classifier once again outperformed all the classifiers and achieved a classification accuracy of 95.6%. SVC remained second with 73.4%; logistic regression was at the third position with 62.7% accuracy. LinearSVC accuracy was 61.4%, while MultinomialNB gave the lowest accuracy of 66.4%.

The results of all the classifiers with lexical features for both datasets, Ubuntu and TripAdvisor, are shown in Figures 2 and 3 , respectively. Experimental observations can be summarized as follows:

(1) With lexical features, the proposed answer detection model outperformed all other classifiers with 95.4% accuracy for Ubuntu and 95.6% for TripAdvisor (NYC) datasets.

(2) For both of the datasets, Ubuntu and TripAdvisor (NYC), lexical features gave almost the same accuracy.

Besides comparison with state-of-the-art classifiers, we also compared the proposed lexical feature-based model for answer detection with the baseline work presented in [8]. In the baseline, an approach was presented to identify the relevant replies to the question in discussion forum threads. Both nonlexical and lexical features were used in this approach. Initially, twenty-eight question-reply relevancy/similarity features were used to retrieve replies that are more relevant to question. In the next step, various feature selection techniques are used to reduce feature space to twelve features. The model improved accuracy for the top twelve

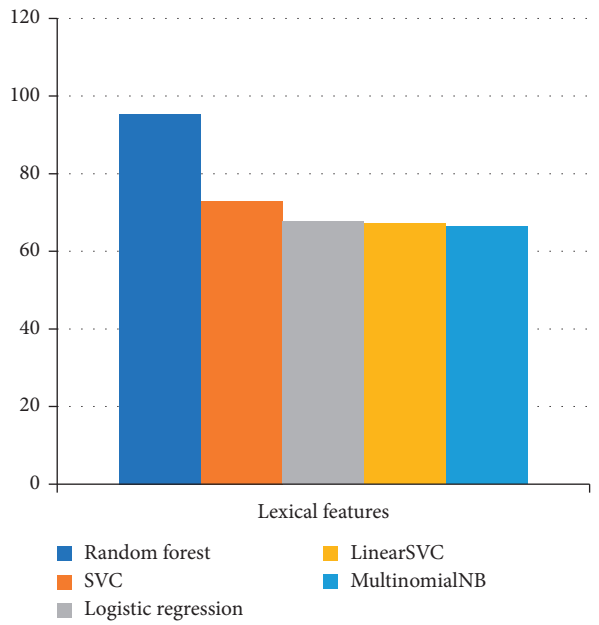

Figure 2: Classifiers accuracy with lexical features for the Ubuntu dataset.

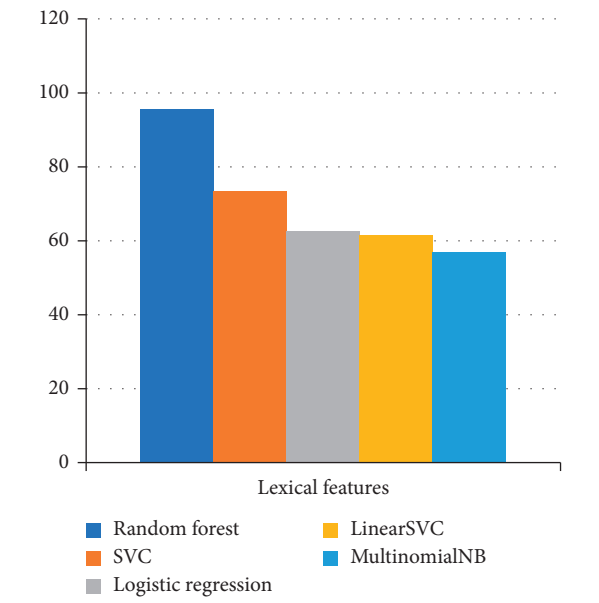

Figure 3: Classifiers accuracy with lexical features for the TripAdvisor (NYC) dataset.

Table 5: Classification accuracy of the proposed model with baseline technique.

Metrics	Baseline	Proposed model	Dataset
Accuracy (%)	79.82	95.4	Ubuntu
Accuracy (%)	76.83	95.6	TripAdvisor (NYC)

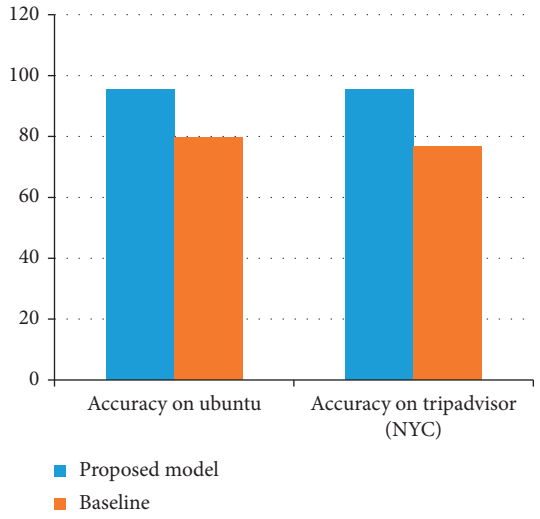

Figure 4: Classification accuracy of the proposed model vs baseline technique for Ubuntu and TripAdvisor (NYC) datasets.

quality dimension features as compared to all twenty-eight features.

The proposed answer detection model has an edge over the baseline work. It is summarized as follows:

The baseline work is based on both lexical and nonlexical features, while the proposed model uses only lexical features. There are no forum-specific features; therefore, it can easily be generalised to any text similarity/relevancy-based problem.

The baseline used top twelve relevancy/similarity features, while the proposed work used only ten features.

The accuracy of baseline technique is 79.82% and 76.83% for Ubuntu and TripAdvisor (NYC) datasets, respectively, while the accuracy of the proposed model for two different forum datasets is 95.4% and 95.6%, as shown in Table 5.

The classification accuracy results for the proposed answer detection model and baseline technique for two forum datasets are visualized in Figure 4.

5. Conclusion and Future Work

The proposed study used lexical features to find similarity and relevancy among two text documents. In this study, we investigated the role of lexical features for detection of relevant answers in online discussion forums. To fairly evaluate the performance of the proposed model, we made a comparison with the baseline technique and other broadly used machine learning models. We performed experiments on two publicly available datasets: Ubuntu and TripAdvisor (NYC). For both datasets, experimental results revealed that the proposed lexical feature-based answer detection model has a greater edge in terms of answer classification accuracy over the baseline technique, which combines both lexical and nonlexical features.

The proposed answer detection model also gave the highest classification accuracy as compared to the other state-of-the art classification model. The proposed model is a supervised model for retrieving answers/replies that are relevant to question/initial post in a discussion forum thread. The model is using a random forest classifier and is totally based on lexical features. The proposed model (based only on lexical features) outperformed the baseline model (based on both lexical and nonlexical features).

For future work, we are planning to extend this work for thread summarization. First, use lexical features and then nonlexical features for thread summarization and then compare the summarization results. We would also like to add some more lexical and nonlexical features and then compare their results.

Data Availability

The data are publicly available at https://ubuntuforums.org and https://www.tripadvisor.com.my/ShowForum-g28953-i4-New_York.html.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the Taif University Researchers Supporting Project (TURSP-2020/231), Taif University, Taif, Saudi Arabia.

References

[1] F. Lin, W. Lei, S.-L. Liu, and G.-C. Liu, "Classification of discussion threads in MOOC forums based on deep learning," *DEStech Transactions on Computer Science and Engineering*, vol. 2017, 2017.

[2] L. Hong and B. D. Davison, "A classification-based approach to question answering in discussion boards," in *Proceedings of Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 171–178, Boston, MA, USA, July 2009.

[3] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media," in *Proceedings of Proceedings of the 2008 International Conference on Web Search and Data Mining*, pp. 183–194, Melbourne, VIC, Australia, February 2008.

[4] H. Hu, B. Liu, B. Wang, M. Liu, and X. Wang, "Multimodal DBN for predicting high-quality answers in cQA portals," in *Proceedings of Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, vol. 2, pp. 843–847Short Papers), Sofia, Bulgaria, August 2013.

[5] B. Liu, J. Feng, M. Liu, H. Hu, and X. Wang, "Predicting the quality of user-generated answers using co-training in

community-based question answering portals," *Pattern Recognition Letters*, vol. 58, pp. 29–34, 2015.

[6] M. Liu, B. Lang, and Z. Gu, "Calculating semantic similarity between academic articles using topic event and ontology," 2017, https://arxiv.org/abs/1710.08011.

[7] M. Farouk, "Measuring sentences similarity: a survey," 2019, https://arxiv.org/abs/1910.03940.

[8] A. Osman, N. Salim, and F. Saeed, "Quality dimensions features for identifying high-quality user replies in text forum threads using classification methods," *PloS One*, vol. 14, 2019.

[9] G. Cong, L. Wang, C.-Y. Lin, Y.-I. Song, and Y. Sun, "Finding question-answer pairs from online forums," in *Proceedings of Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 467–474, Singapore, July 2008.

[10] G. Zhou, Y. Liu, F. Liu, D. Zeng, and J. Zhao, "Improving question retrieval in community question answering using world knowledge," in *Proceedings of Twenty-Third International Joint Conference on Artificial Intelligence*, Beijing, China, August 2013.

[11] V. S. Shirsat, R. S. Jagdale, and S. N. Deshmukh, "Sentence level sentiment identification and calculation from news articles using machine learning techniques," in *Proceedings of the Computing, Communication and Signal Processing*, pp. 371–376, Springer, Dalian, China, September 2019.

[12] P. Biyani, S. Bhatia, C. Caragea, and P. Mitra, "Using non-lexical features for identifying factual and opinionative threads in online forums," *Knowledge-Based Systems*, vol. 69, pp. 170–178, 2014.

[13] R. C. Kanjirathinkal, A. Singh, R. Gangadharaiah, D. Raghu, and K. Visweswariah, "Does similarity matter? The case of answer extraction from technical discussion forums," in *Proceedings of the COLING Posters*, pp. 175–184, Mumbai, India, December 2012.

[14] K. Chai, C. Wu, V. Potdar, and P. Hayati, "Automatically measuring the quality of user generated content in forums," in *Proceedings of the Australasian Joint Conference on Artificial Intelligence*, pp. 51–60, Canberra Australia, December 2020.

[15] A. I. Obasa, N. Salim, and A. Khan, "Enhanced lexicon based model for web forum answer detection," in *Proceedings of 2015 Fifth International Conference on Digital Information Processing and Communications (ICDIPC)*, pp. 237–243, Sierre, Switzerland, October 2015.

[16] D. Ö. Şahin and E. Kılıç, "Two new feature selection metrics for text classification," *Automatika*, vol. 60, pp. 162–171, 2019.

[17] S. Dey Sarkar, S. Goswami, A. Agarwal, and J. Aktar, "A novel feature selection technique for text classification using Naive Bayes," *International Scholarly Research Notices*, vol. 2014, Article ID 717092, 10 pages, 2014.

[18] J. Jeon, W. B. Croft, J. H. Lee, and S. Park, "A framework to predict the quality of answers with non-textual features," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 228–235, Berkeley, CA, USA, August 1999.

[19] J. Huang, M. Zhou, and D. Yang, "Extracting chatbot knowledge from online discussion forums," in *Proceedings of the IJCAI*, pp. 423–428, Tokyo, Japan, August 1979.

[20] K. Soumya George and S. Joseph, "Text classification by augmenting bag of words (BOW) representation with co-occurrence feature," *IOSR Journal of Computer Engineering*, vol. 16, pp. 34–38, 2014.

[21] A. I. Obasa, N. Salim, and A. Khan, "Hybridization of bag-of-words and forum metadata for web forum question post detection," *Indian Journal of Science and Technology*, vol. 8, pp. 1–12, 2016.

[22] L. Yang, C. Li, Q. Ding, and L. Li, "Combining lexical and semantic features for short text classification," *Procedia Computer Science*, vol. 22, pp. 78–86, 2013.

[23] A.-S. Mohammad, Z. Jaradat, A.-A. Mahmoud, and Y. Jararweh, "Paraphrase identification and semantic text similarity analysis in Arabic news tweets using lexical, syntactic, and semantic features," *Information Processing & Management*, vol. 53, pp. 640–652, 2017.

[24] Y. Liu, C.-J. Sun, L. Lin, X. Wang, and Y. Zhao, "Computing semantic text similarity using rich features," in *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pp. 44–52, Shanghai, China, October 2015.

[25] A. Islam and D. Inkpen, "Semantic text similarity using corpus-based word similarity and string similarity," *ACM Transactions on Knowledge Discovery from Data*, vol. 2, no. 2, pp. 1–25, 2008.

[26] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 649–657, Lake Tahoe, NV, USA, December 2012.

[27] X. Wu, V. Kumar, J. Ross Quinlan et al., "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, 2008.

[28] V. Kharde and P. Sonawane, "Sentiment analysis of twitter data: a survey of techniques," 2016, https://arxiv.org/abs/1601.06971.

[29] S. Scott and S. Matwin, "Text classification using WordNet hypernyms," in *Proceedings of the Usage of WordNet in Natural Language Processing Systems*, Quebec, CA, USA, August 1998.