Hindawi
Scientific Programming
Volume 2020, Article ID 4621196, 10 pages
https://doi.org/10.1155/2020/4621196

*Research Article*

# Machine Learning Approach for Answer Detection in Discussion Forums: An Application of Big Data Analytics

**Atif Khan,[1] Ibrahim Ibrahim,[1] M. Irfan Uddin,[2] Muhammad Zubair,[1] Shafiq Ahmad,[3] Muhammad Dzulqarnain Al Firdausi,[3] and Mazen Zaindin[4]**

[1]*Department of Computer Science, Islamia College Peshawar, Peshawar, Pakistan*
[2]*Institute of Computing, Kohat University of Science and Technology, Kohat, Pakistan*
[3]*King Saud University, College of Engineering, Department of Industrial Engineering, Riyadh, Saudi Arabia*
[4]*King Saud University, College of Science, Department of Statistics and Operations Research, Riyadh, Saudi Arabia*

Correspondence should be addressed to Shafiq Ahmad; ashafiq@ksu.edu.sa

Received 26 February 2020; Accepted 15 April 2020; Published 22 May 2020

Academic Editor: Shaukat Ali



Nowadays, data are flooding into online web forums, and it is highly desirable to turn gigantic amount of data into actionable knowledge. Online web forums have become an integral part of the web and are main sources of knowledge. People use this platform to post their questions and get answers from other forum members. Usually, an initial post (question) gets more than one reply posts (answers) that make it difficult for a user to scan all of them for most relevant and quality answer. Thus, how to automatically extract the most relevant answer for a question within a thread is an important issue. In this research, we treat the task of answer extraction as classification problem. A reply post can be classified as relevant, partially relevant, or irrelevant to the initial post. To find the relevancy/similarity of a reply to the question, both lexical and nonlexical features are used. We proposed to use LinearSVC, a variant of support vector machine (SVM), for answer classification. Two selection techniques such as chi-square and univariate are employed to reduce the feature space size. The experimental results showed that LinearSVC classifier outperformed the other state-of-the-art classifiers in the context of classification accuracy for both Ubuntu and TripAdvisor (NYC) discussion forum datasets.

## 1. Introduction

Web forum is an online discussion board where like-minded people gather and discuss issues on specific topics. Web forum has become an integral part of the web due to its constant growth. There is a good chance to get forum pages while searching for a question/topic.

Forum users share information on different topics. Discussion among the users is started when one user asks a question and other users/members answer it, so this forms a forum thread where a question gets more than one answers [1].

Question post in a forum thread usually get answers with different qualities. Quality means to what extent a reply post addresses the question. Each user answers the question according to their own understanding and knowledge, which may be relevant, partially relevant, or irrelevant. This makes it difficult for the question poster to identify the most relevant answer [1]. It is very tedious and laborious to go through all the reply posts and then identify the relevant answer. So, the main objective of this research is to automatically extract/identify the most relevant answers/replies for a posted question within a thread.

We consider the thread's initial post as question and all other replies as candidate answers of different qualities. To keep the process simple, we ignore all the questions in the reply posts and topic drift in the thread.

There are two types of features, lexical and nonlexical. Both of them are used to find reply relevancy and similarity with the given question [1–5]. Some nonlexical features are not always available [6], which cannot be calculated easily and also make the model forum dependent, e.g., if forum metadata are used for training the model, then the model

becomes dependent on those specific features, and hence it cannot be easily adapted to other forums. Therefore, in this study, we mostly exploited lexical, content-based, and semantic features to make the model forum independent which can be easily adapted to other forums.

Like other research works [7–9], we also consider answer extraction as text classification problem. Replies are classified into three classes: high-quality, low-quality, and nonquality, depending on their relevancy with the question post. For answer detection/classification, we used support vector machines (SVMs). It is a group of algorithms used for classification, regression, and outlier detection. Two variants, LinearSVC and SVC, of SVM are used here. LinearSVC outperformed the other classifiers and gave high accuracy of 76.3%.

For lexical similarity features, like cosine similarity, we use bag-of-words (BoW) approach to convert text into vectors [10]. Since all features/words are not equally important, redundant ones are devalued by using TfidfVectorizer. In this study, we used unigram, bigram, and trigram word sequence.

Mining for best reply posts within a thread has many applications. Question/answer forums such as *Yahoo! Answers* can suggest answers extracted from forum thread to their users.

It can also be used to generate question-answer pairs which can be further filtered to frequently asked questions (FAQ). Contributions of our work are summarized as follows:

(a) To propose the answer detection model based on support vector machine (SVM) using lexical and nonlexical features.

(b) To enhance the proposed model by identifying optimal feature combination using univariate and chi-square feature selection techniques.

(c) To improve the proposed model by proposing some new semantic features.

The remainder of the paper is organized as follows. Section 2 is about related work. Section 3 explains our proposed framework. Section 4 describes the experimental settings, results, and discussion. Finally, Section 5 concludes and presents the future work.

## 2. Related Work

Predicting answer quality in online web forums is a text classification problem [7–9, 11, 12]. Different approaches and methodologies have been used for this task. Bag-of-words (BoW) approach is a commonly used approach [1]. In this approach, the text is represented by its words and each word is considered as a feature. Frequency of each feature is recorded and a vector is created, which is further used to find the similarity with other vectors. Usually, BoW is used with bigram and trigram to get more information. This approach was augmented with co-occurrence feature from Wikipedia and was used to classify news articles in one of the twenty groups [13]. The authors in [14] integrated BoW approach with forum metadata, simple rule of question mark, and question words to extract questions from web forums. Multimodal deep belief net was used in [8] to check answer quality. This model solved the issue of nonlinear correlation between lexical and nonlexical features. A framework based on convolutional neural network was developed in [11] to classify massive open online course (MOOC) forum threads. Others used character-level ConvNet for text classification [15].

To classify text in web forums as question or nonquestion, a sequential model [2] was proposed, which is based on patterns extracted from questions and nonquestions. The model then used a graph-based approach for answer extraction in the same thread.

Another approach called cascaded framework was introduced in [16] for <thread-title, reply> pair extraction from web forums to enrich chatbot knowledge. In the first step, replies were extracted which were logically relevant to the thread title. Then, the extracted pairs were ranked and top N were selected.

Both types of algorithms, traditional such as Naïve Bayes and deep learning like convolutional neural networks and multimodal deep net, have been used for extracting quality contents from the web forums [1, 8, 11, 15, 17].

Text classification task is based on the quality of contents. Quality means to what extent it is relevant and addressing the query. So, for classification, it is necessary to measure the quality of contents using different features [18]. Reply post in a forum thread is classified as high-quality, low-quality, and nonquality based on their relevancy with the question post.

Primarily, there are two types of features, lexical and nonlexical, used for answer extraction within a thread. These are categorized in different ways: the authors in [1] identified six feature groups and further divided them into 28 subfeatures. The authors in [6] described five types of features that are lexical, content base, structural, forum specific, and reply-to and further divided them into 17 subfeatures.

In some forums, lexical similarity cannot be used much effectively because answers have very minimal overlap with the questions and nonanswers also show the same behaviour [6]. In such cases, nonlexical features are more reliable than lexical ones. In some cases, researchers proposed a framework totally relying on nonlexical features for judging the quality of documents [19]. Some researchers showed that combining n-gram of lexical with nonlexical features gave good results [7]. The authors in [11] used *user interactive behaviour* features to classify massive open online course (MOOC) threads using convolutional neural network, as the model based on such features are language and content independent. The authors in [16] used structural and content-based features to develop their framework for <title, reply> pair extraction to enrich chatbot knowledge. The authors in [5] used nonlexical thread features to classify web forum threads into subjective and nonsubjective. Thus, different research studies used various combinations of features to enhance the model performance. One such study identified 12 features while the other identified 6 best features [1, 6].

In a nutshell, not all features are important; some do not contribute while others negatively affect the model

performance, so in order to get optimal subfeature list, the authors in [1, 20] eliminated nonvaluable and redundant features. Moreover, forum noise also adversely affects the model performance [20]. On the other hand, normalizing the forum noise will enhance the model performance. Hence, how to select best features list is nontrivial due to different nature of forum data.

There are different selection techniques to reduce feature space size. Mainly, these are grouped into filtered, wrapper, and embedded methods. *Document frequency thresholding (DF), chi-squared (CHI), information gain (IG)*, and *Acc and Acc2 (Acc2)* are the most commonly used feature selection techniques [21]. The authors in [17] used univariate and clustering feature techniques to improve the Naïve Bayes performance for text classification task. Authors in [21] have introduced two new feature selection metrics for text classification such as Relevance Frequency Feature Selection (RFFS) and Alternative Accuracy2 (AAcc2); and suggested that the new metrics produced promising results as the current frequenly used metrics. Other researchers used information gain (IG), chi-square, and gain ratio (GR) to get top 12 best features. To gauge the significance of each feature, permutation and ablation tests are also performed [6].

However, our proposed study used LinearSVC to classify reply posts within a forum thread. For feature space size reduction, univariate and chi-square selection techniques are used to select optimal subfeature list. The next section describes our proposed methodology.

## 3. Proposed Methodology

Our proposed model is summarized in Figure 1. It is divided into four phases: in the first phase, data are preprocessed to eliminate errors and noise. In the second phase, lexical and nonlexical features for the question and reply posts are calculated to find their similarities. Thirdly, features are filtered using different selection techniques. In the final phase, the kernel method of SVM called LinearSVC is used to classify the replies as high-quality, low-quality, and nonquality. These steps are explained below.

*3.1. Preprocessing.* Converting raw data into predictable and analyzable format is data preprocessing. The following steps are taken to preprocess the data:

(a) Converting all words to lowercase

(b) Lemmatizing words using WordNetLemmatizer of NLTK

(c) Removing all stop words

(d) Expanding the abbreviation

*3.2. Feature Extraction.* Different features are used to find the relevancy and similarity of a reply post to its initial post. These features are categorized in different ways. A study conducted by Osman et al. [1] categorized features into six groups that are relevancy, author activeness, timeliness, ease-of-understanding, amount-of-data, and politeness. These groups were further divided into 28 subfeatures. Similarly, another research study identified five feature groups: lexical, content, forum specific, structural, and reply-to types, and further divided them into 17 subfeatures [6].

Broadly, features are classified into lexical and nonlexical. Lexical features are text-specific features, e.g., cosine similarity of question and reply posts. Similarly, the number of unique words in a reply post is also a lexical feature. Nonlexical features are forum-specific (author or thread structure related) and content-based features. Total number of threads the users have participated, author reputation in the forum, and time elapsed between question and reply posts are some examples of nonlexical features.

For answer extraction, in discussion forums, some researchers have preferred nonlexical features over lexical ones [5–7, 19], while others have proposed lexical features [20]. Naturally, questions have some kind of lexical similarity with their answers [20], so one should use both lexical and nonlexical features to extract most relevant and quality answer [8]. Lexical features are used to find the relevancy of answer with the question, while nonlexical features are used to check their quality [19] that is to what extent an answer addresses the question.

Some features are not always available. One researcher inspected 12 data forums and found that 36.3% forum-specific features are available, while 75% author activeness features were available [6]. In our case, timeliness features are not available. Moreover, using some features makes the model forum specific. So, in this study, we used both lexical and nonlexical features particularly targeting those features which are 100% available and can be easily calculated from the text or structure of the thread. These features are lexical, content-based, and semantic features.

In this study, we used twenty features given in Table 1 with brief description. Out of these, fourteen features are lexical, content-based, or semantic features as shown in Table 2. In the table, the three highlighted features F1, F16, and F17 are our new proposed semantic features. Some features like F7, F11, F12, F13, and F20 are directly calculated from the text or thread structures. For example, F7 is the number of unique words in a reply post which can be calculated by splitting it in words and then applying *set* and *len* functions in Python Language.

For pure lexical features like F2, F3, F4, F5, and F6, we used bag-of-words (BoW) approach. BoW approach is a well-known technique to extract features from documents and represent them as vector. Vector values represent number of occurrences of a word in the documents. Since BoW approach ignores feature order and only word frequency matters, to preserve sentence structure and words order, we used bigram and trigram word sequences which will get more meaning from the document. Some features get high frequency but are not much valuable, so for filtering unimportant features/words, we used the term frequency inverse document frequency (TF-IDF) technique, which converts text into vectors and assigns weightage to each word according to their importance in the document.

We introduced three new semantic features called F1, F16, and F17 for answer extraction in discussion forums, and to the best of our knowledge, these features have not been used in the literature. We used word mover distance and

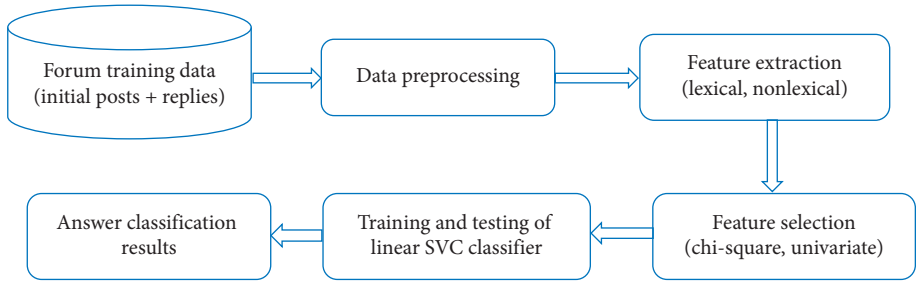


Figure 1: Proposed approach for answer detection in discussion forums.

Table 1: All twenty features with brief description.

| Code | Abbreviation | Description | Feature type | Subtype |
|---|---|---|---|---|
| F1 | ThrdCntrodRplyWMDistance | Word mover distance of a reply from the thread centre | Lexical | Semantic |
| F2 | ThrdCentrodRplyCosnSmlrty | Cosine similarity of a reply with the thread centre | Lexical | Pure lexical |
| F3 | TtlRplyCosnSmlrtyWholCrps | Cosine similarity of a reply with the title based on corpus created from all threads | Lexical | Pure lexical |
| F4 | QustionRplyCosnSmlrtyWholCrps | Cosine similarity of a reply with the initial post based on corpus created from all threads | Lexical | Pure lexical |
| F5 | TtlRplyCosnSmlrty | Cosine similarity of a reply with the thread title | Lexical | Pure lexical |
| F6 | QustionRplyCosnSmlrty | Cosine similarity of a reply with the thread initial post | Lexical | Pure lexical |
| F7 | UnqWrds | Number of unique words in a reply | Lexical | Pure lexical |
| F8 | IsRplyByCrtrOfInitlPost | Was the reply given by the creator of initial post? | Nonlexical | Structural |
| F9 | NumRepliesByUsrCurrentThrd | Total number of replies given by the user in the current thread | Nonlexical | Structural |
| F10 | NoThrdsUsrParticipated | Total number of threads the user has participated | Nonlexical | Structural |
| F11 | ReWrdsOvrlpInitialPost | Number of overlapping words between the initial post and the reply post | Lexical | Pure lexical |
| F12 | ReWrdsOvrlpThrdTitl | Number of overlapping words between the thread title and the reply post | Lexical | Pure lexical |
| F13 | IsRplyContan5WHWrds | Does the reply contain 5WH words? | Nonlexical | Content based |
| F14 | IsRplyMntionOthrUsrNames | Does the reply refer to any other forum user? | Nonlexical | Structural |
| F15 | IsRplyHvHyperlnk | Does the reply have any Hyperlink? | Nonlexical | Content based |
| F16 | WMDbtwnTitlRpl | Word mover distance between thread title and reply | Lexical | Semantic |
| F17 | WMDbtwnQustionRpl | Word mover distance between initial post and reply | Lexical | Semantic |
| F18 | TotlNoRpliesByUsrInAllThrds | Total number of replies given by the user in all threads | Nonlexical | Structural |
| F19 | TotlNoIntialPstsByUser | Total number of initial posts created by the user | Nonlexical | Structural |
| F20 | NoWrdsRply | Total number of words present in a reply | Lexical | Pure lexical |

Google's pretrained word2vec model for our proposed new features. Google's pretrained word2vec model, used for contextual/semantic similarity of words, has vectors for three million words/phrases, and it has been trained on roughly hundred billion words from Google News dataset. We leave the default word vector length to be 300 features and hence the word2vec model will check the relevance of two words in 300 dimensional space. Its speciality is words having same semantic/context will have close vectors. Word mover (WM) distance is the measure of dissimilarity of two documents. The greater the WM distance, the greater will be the dissimilarity and vice versa. Zero distance means that the two documents are completely related with each other.

Feature F1 is the contextual similarity of each reply with the thread centroid. For thread centroid, the most important features/words are obtained using the TF-IDF technique. Word mover distance of each reply from the centroid is calculated using Google's pretrained word2vec model. Feature F16 is the word mover distance of thread title and a reply while feature F17 is the word mover distance of the initial post/question and a reply.

The proposed new semantic features (F1, F16, and F17) are the important ones since both chi-square and univariate feature selection techniques selected them in the top features space for both Ubuntu and TripAdvisor(NYC) datasets as given Tables 3–6.

*3.3. Feature Selection.* There is a list of features, lexical and nonlexical, which can be used for extracting answer in the question-answer forums. But all of them are not equally important and cannot be used due to the following reasons:

(a) Some features are nonvaluable and negatively affect the model performance [1]

Table 2: Fourteen lexical, content-based, and semantic features (including proposed semantic features F1, F16, and F17).

| Code | Abbreviation |
|---|---|
| F1 | ThrdCntrodRplyWMDistance |
| F2 | ThrdCentrodRplyCosnSmlrty |
| F3 | TtlRplyCosnSmlrtyWholCrps |
| F4 | QustionRplyCosnSmlrtyWholCrps |
| F5 | TtlRplyCosnSmlrty |
| F6 | QustionRplyCosnSmlrty |
| F7 | UnqWrds |
| F11 | ReWrdsOvrlpInitialPost |
| F12 | ReWrdsOvrlpThrdTitl |
| F13 | IsRplyContan5WHWrds |
| F15 | IsRplyHvHyperlnk |
| F16 | "WMDbtwnTitlRpl" |
| F17 | "WMDbtwnQustionRpl" |
| F20 | NoWrdsRply |

Table 3: Top 11 features selected by the chi-square technique for Ubuntu dataset.

| Code | Abbreviation |
|---|---|
| F1 | ThrdCntrodRplyWMDistance |
| F2 | ThrdCentrodRplyCosnSmlrty |
| F7 | UnqWrds |
| F8 | IsRplyByCrtrOfInitlPost |
| F9 | NumRepliesByUsrCurrentThrd |
| F13 | IsRplyContan5WHWrds |
| F15 | IsRplyHvHyperlnk |
| F16 | "WMDbtwnTitlRpl" |
| F17 | "WMDbtwnQustionRpl" |
| F19 | TotlNoIntialPstsByUser |
| F20 | NoWrdsRply |

Table 4: Top 15 features selected by the univariate technique for Ubuntu dataset.

| Code | Abbreviation |
|---|---|
| F1 | ThrdCntrodRplyWMDistance |
| F2 | ThrdCentrodRplyCosnSmlrty |
| F3 | TtlRplyCosnSmlrtyWholCrps |
| F5 | TtlRplyCosnSmlrty |
| F6 | QustionRplyCosnSmlrty |
| F7 | UnqWrds |
| F8 | IsRplyByCrtrOfInitlPost |
| F10 | NoThrdsUsrParticipated |
| F11 | ReWrdsOvrlpInitialPost |
| F13 | IsRplyContan5WHWrds |
| F15 | IsRplyHvHyperlnk |
| F16 | "WMDbtwnTitlRpl" |
| F17 | "WMDbtwnQustionRpl" |
| F19 | TotlNoIntialPstsByUser |
| F20 | NoWrdsRply |

(b) Some features are correlated while some are obtained from other feature combination

(c) Not all features are available in datasets

Table 5: Top 8 features selected by the chi-square technique for TripAdvisor (NYC) dataset.

| Code | Abbreviation |
|---|---|
| F2 | ThrdCentrodRplyCosnSmlrty |
| F6 | QustionRplyCosnSmlrty |
| F7 | UnqWrds |
| F8 | IsRplyByCrtrOfInitlPost |
| F13 | IsRplyContan5WHWrds |
| F17 | WMDbtwnQustionRpl |
| F19 | TotlNoIntialPstsByUser |
| F20 | NoWrdsRply |

Table 6: Top 10 features selected by the univariate technique for TripAdvisor (NYC) dataset.

| Code | Abbreviation |
|---|---|
| F1 | ThrdCntrodRplyWMDistance |
| F2 | ThrdCentrodRplyCosnSmlrty |
| F6 | QustionRplyCosnSmlrty |
| F7 | UnqWrds |
| F8 | IsRplyByCrtrOfInitlPost |
| F9 | NumRepliesByUsrCurrentThrd |
| F16 | WMDbtwnTitlRpl |
| F17 | WMDbtwnQustionRpl |
| F19 | TotlNoIntialPstsByUser |
| F20 | NoWrdsRply |

(d) Using forum-specific features makes the model forum dependent

(e) Using all of them is computationally expensive

To overcome the above limitations, initially we select those features whose availability is hundred percent and can be easily calculated from the text as discussed in Section 3.2. Then, we employed two feature selection techniques, namely, chi-square and univariate, to reduce the feature space size in order to get optimal features as discussed in detail in Section 4.3.

*3.4. Classification Model Construction.* This phase aims to classify the reply posts as relevant, partially relevant, and irrelevant using machine learning algorithm. We used a kernel method of support vector machine (SVM) called LinearSVC. This classification is based on the relevancy of a reply to the initial post.

We compared the classification accuracy of the LinearSVC classifier with other kernel methods of SVM as well as other state-of-the-art classification algorithms such as multinomial Naïve Bayes, Bernoulli Naïve Bayes, random forest, and logistic regression. All classifiers were trained and tested with three sets of features that are all features and two subfeature sets chosen by different feature selection techniques. More details can be found in Section 4.

## 4. Experimental Settings

*4.1. Evaluation Data.* The proposed answer detection model is evaluated on two datasets—the online TripAdvisor forum

(https://www.tripadvisor.com.my/ShowForum-g28953-i4-New_York.html) for New York City (NYC) and online Ubuntu Linux distribution forum (http://ubuntuforums.org). The authors randomly chose 100 threads from both forums, each having a question and multiple replies. There are total 756 replies in Ubuntu and 788 replies in TripAdvisor (NYC) dataset. Replies have been categorized into three classes. Reply which is completely relevant is assigned a class label 3, partially relevant reply is assigned a class label 2, and 1 is assigned to irrelevant replies. . Both of the datasets have 7 columns, "ThreadID," "Title," "UserID_inipst," "Questions," "UserID," "Replies," and "Class," for each thread. We split the labelled dataset in such a way that 80% data is used for training and 20% data is used for testing.

*4.2. Classification Algorithms.* We chose a linear kernel method of support vector machine (SVM) called LinearSVC for classification of answer/reply post in text forum threads. SVM is widely used for text classification problem [22]. We also compared the performance of LinearSVC with other kernel methods of SVM as well as other state-of-the-art classification algorithms. The classifiers are briefly discussed as follows.

*4.2.1. Naïve Bayes.* It is a group of supervised learning algorithms based on Bayes theorem which considers each feature as independent of other features. This classifier has been largely used in text classification problems and has given good results [23].

Bayes theorem is stated below:

$$P(y \mid x_1, x_2, ...., x_n) = \frac{P(y)P(x_1, x_2, ...., x_n \mid y)}{P(x_1, x_2, ...., x_n)}, \quad (1)$$

where $y$ is class variable and $x_1$ to $x_n$ represents a dependent feature vector.

Naïve Bayes requires small amount of data to train and is extremely fast compared to other classifiers.

The following variant of Naïve Bayes is used in the evaluation of this study.

Multinomial Naïve Bayes It is used for multinomial distributed data. It is mainly used for text classification.

*4.2.2. Support Vector Machines (SVMs).* It is a group of algorithms used for classification, regression, and outlier detection. It performs well in high dimensional space [4] and uses less memory. It works with different kernels, and custom kernel can also be specified. We used the following three implementations.

*Support Vector Classification (SVC).* It is based on libsvm. Its fit time increases quadratically with the number of samples. "rbf" is the default kernel. Other kernels are "linear," "poly," and "sigmoid."

*NuSVC.* It is the same as SVC but has slightly different parameter set and mathematical formulation. It is based on *libsvm*. Here, *Nu* is a regularization parameter having values from 0 to 1. The parameters *C* and *Nu* are same in the context of their classification power but evaluation of *Nu* is easier than *C*.

*LinearSVC.* It is based on "liblinear" with "linear" kernel. Input could be dense or sparse and is more flexible in choosing of penalties and loss functions.

*4.2.3. Logistic Regression.* It is a classification method that generalizes logistic regression (LR) to multiclass problems, i.e., more than 2 discrete outcomes. It is a model used to predict probabilities of different outcomes of a target variable given a set of input features.

*4.2.4. Random Forests.* They are also known as random decision forests. They are an ensemble learning method for classification task and work by building a large number of decision trees at training time and output the class that is the mode of the classes (classification) of the individual decision trees.

*4.3. Feature Reduction.* To eliminate nonvaluable and redundant features, two selection techniques are employed: chi-square and univariate. The former selected eleven best features for Ubuntu and eight best features for TripAdvisor dataset shown in Tables 3 and 5, respectively, while the latter one selected fifteen optimal features for Ubuntu and ten best features for TripAdvisor dataset shown in Tables 4 and 6, respectively. In the following section, it has been shown that classifiers with these subsets of features performed well than those with all features.

*4.4. Experimental Results and Discussion.* Results of all six classifiers, used in this study, with all features and features selected by different selection techniques are discussed in this section. LinearSVC, SVC, NuSVC, MultinomialNB, random forest (RF), and logistic regression (LR) are used in this study.

In the first phase, classifiers were used for Ubuntu dataset for all twenty features as shown in Table 7. All the six classifiers gave good results, but MultinomialNB and LinearSVC performed well and gave exactly the same accuracy of 73.7%. LR has the second highest accuracy (72.4%) while SVC resulted in 71.1% accuracy. Random Forest occupied fourth position with 63.2% accuracy.

Then, classifiers were tested for TripAdvisor dataset using all twenty features. The results are shown in Table 8. It is clear from the results that LinearSVC has the highest accuracy of 68.4%. RF and LR are at the second position with 67.1% accuracy while NuSVC has 64.6% accuracy. SVC and MultinomialNB have the lowest accuracy.

In the second phase, feature space was reduced by employing the chi-square feature selection technique. Eleven best features were selected for Ubuntu and eight best features were chosen for TripAdvisor dataset (Tables 3 and 5). The three new semantic features, introduced in this work, were picked by the feature selection technique. This shows that

Table 7: Results for Ubuntu dataset using all features.

| Classifiers | Accuracy (%) |
|---|---|
| LinearSVC | *73.7* |
| SVC | 71.1 |
| MultinomialNB | **73.7** |
| Random forest | 63.2 |
| Logistic regression | 72.4 |
| NuSVC | 61 |

Table 8: Results for TripAdvisor dataset using all features.

| Classifiers | Accuracy (%) |
|---|---|
| LinearSVC | **68.4** |
| NuSVC | 64.6 |
| Random forest | 67.1 |
| Logistic regression | 67.1 |
| SVC | 62.5 |
| MultinomialNB | 59 |

Table 9: Results for Ubuntu dataset using top 11 features selected by the chi-square technique.

| Classifiers | Accuracy (%) |
|---|---|
| LinearSVC | **73.7** |
| SVC | 67.1 |
| MultinomialNB | 72.4 |
| Logistic regression | 72.4 |
| Random forest | 64 |
| NuSVC | 63.1 |

Table 10: Results for TripAdvisor dataset using top 8 features selected by the chi-square technique.

| Classifier | Accuracy (%) |
|---|---|
| LinearSVC | **76** |
| NuSVC | 67.1 |
| Random forest | 65.8 |
| Logistic regression | 73.4 |
| SVC | 64 |
| MultinomialNB | 62 |

these features are the important ones for question-reply similarity.

Classifiers were employed for Ubuntu dataset with these optimal features. The results in Table 9 show again that LinearSVC has the highest accuracy of 73.7%. MultinomialNB and LR have the same accuracy of 72.4%. SVC is at fourth position with 67.1% accuracy. Random forest is at fifth and NuSVC is at sixth position. Referring to Tables 7 and 9, LinearSVC and LR gave the same accuracy as those with the twenty features. Random forest and NuSVC also increased their accuracy. MultinomialNB's accuracy was slightly reduced, but this time only 11 features were used instead of twenty.

Results of all six classifiers, LinearSVC, NuSVC, RF, LR, SVC, and MultinomialNB, for TripAdvisor dataset with top eight best features selected by the chi-square technique are shown in Table 10. Again, LinearSVC performed well with 76% accuracy while LR is at second position and NuSVC is at third position with 73.4% and 67.1% accuracies, respectively. RF has the lowest accuracy (65.8%).

LinearSVC's accuracy increased by 7.6%, NuSVC's accuracy increased by 2.5%, and LR's accuracy increased by 6.3% as compared to the accuracy with all twenty features given in Table 8. SVC and MultinomialNB also increased their accuracy.

In the third phase, the univariate feature selection technique was employed to filter the features. Fifteen best features were selected for Ubuntu dataset and ten were selected for TripAdvisor dataset, as shown in Tables 4 and 6, respectively. Again, the newly introduced three semantic features were also selected for both of the datasets.

Classifier results, with the selected features, for Ubuntu dataset are given in Table 11. LinearSVC is at top with 76.3% accuracy. MultinomialNB's accuracy is 73.7%. SVC and LR

Table 11: Results for Ubuntu dataset using top 15 features selected by the univariate technique.

| Classifiers | Accuracy (%) |
|---|---|
| MultinomialNB | 73.7 |
| LinearSVC | **76.3** |
| SVC | 72.4 |
| Random forest | 60.5 |
| Logistic regression | 72.4 |
| NuSVC | 60 |

Table 12: Results for TripAdvisor dataset using top 10 features selected by the univariate technique.

| Classifiers | Accuracy (%) |
|---|---|
| LinearSVC | **73.4** |
| SVC | 69.6 |
| NuSVC | 67.1 |
| Random forest | 69.6 |
| Logistic regression | 72.2 |
| MultinomialNB | 61.5 |

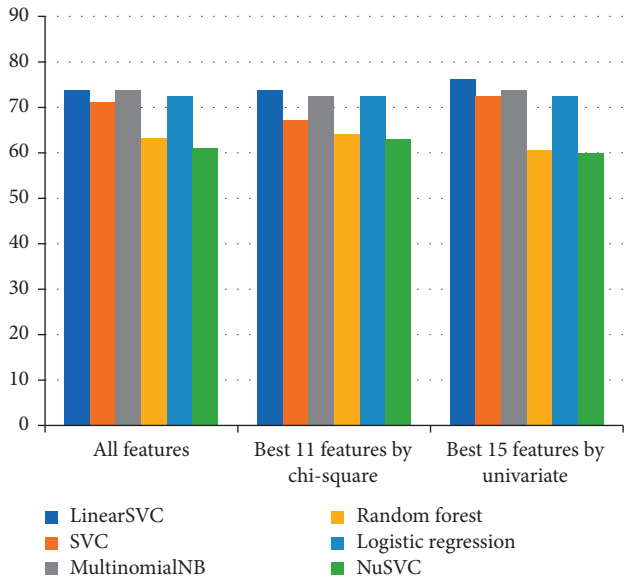


Figure 2: Classifiers' accuracy based on different feature sets for Ubuntu dataset.

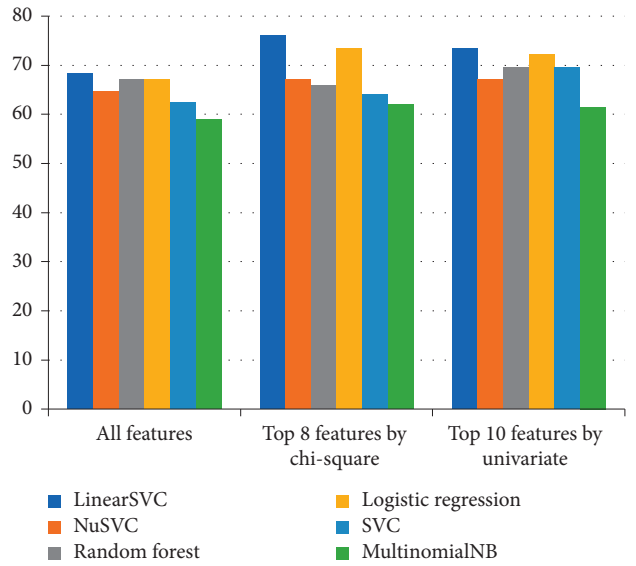


Figure 3: Classifiers' accuracy based on different feature sets for TripAdvisor dataset.

have the same accuracy of 72.4% while RF's accuracy is 60.5%. The classifiers performed better, with the selected fifteen features, as compared to all twenty features.

For TripAdvisor dataset, algorithms were used with top ten selected best features chosen by the univariate feature selection technique. Results in Table 12 show that the classifiers' performance is much better than that with all twenty features. LinearSVC's accuracy increased from 68.4% to 73.4%. The accuracy of NuSVC and RF was increased by 2.5% while LR improved its accuracy from 67.1% to 72.2%.

The classification accuracy of different classifiers based on different features in the context of Ubuntu and TripAdvisor datasets are depicted in Figures 2 and 3, respectively. From the experimental results, we observed the following:

(1) Most of the classifiers' accuracy increased or remained unchanged with best selected features.

(2) Our proposed classifier *LinearSVC* outperformed all other state-of-the-art classifiers.

(3) Our new proposed three semantic features were selected by the selection techniques for both of the datasets and greatly improved the accuracy of LinearSVC classifier.

## 5. Conclusion and Future Work

Automatic solution for extracting most relevant and quality answer to the initial post (question) in the thread/discussion forum is a challenging task. This study sets a new direction

by presenting lexical, content-based, and semantic features that greatly improved the classification accuracy of the proposed classifier. In this study, we proposed to use a supervised machine learning model for extracting most relevant replies to the initial post, within a forum thread, using a kernel method of support vector machine called LinearSVC and compared it with other SVM kernel methods and other state-of-the-art classification algorithms. LinearSVC, a variant of SVM, gave highest accuracy. Two subsets of features were explored, which improved the model performance. Moreover, three new semantic features were introduced and selected as best features by both chi-square and univariate feature selection techniques which significantly improved the accuracy of LinearSVC. For Ubuntu dataset, the chi-square technique selected 6 lexical and 5 nonlexical features, while the univariate technique selected 10 lexical and 5 nonlexical features. For TripAdvisor (NYC), the chi-square technique selected 5 lexical and 3 nonlexical features while the univariate technique selected 7 lexical and 3 nonlexical features. So, lexical features proved more imperative and vital for answer extraction in discussion boards.

In future, we plan to explore more semantic and content-based features to further enhance the model. Also, this work can be extended to thread summarization.

## Data Availability

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

## Conflicts of Interest

The authors declare no conflicts of interest.

## Acknowledgments

The authors extend their appreciation to the Deanship of Scientific Research at King Saud University for funding this work through research group no. RG-1438-089.

## References

[1] 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, no. 5, 2019.

[2] G. Cong, "Finding question-answer pairs from online forums," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pulau Ujong, Singapore, July 2008.

[3] G. Zhou et al., "Improving question retrieval in community question answering using world knowledge," in *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, Beijing, China, August 2013.

[4] V. S. Shirsat, R. S. Jagdale, and S. N. Deshmukh, "Sentence level sentiment identification and calculation from news articles using machine learning techniques," in *Computing, Communication and Signal Processing*, pp. 371–376, Springer, Berlin, Germany, 2019.

[5] 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.

[6] R. C. Kanjirathinkal, "Does similarity matter? The case of answer extraction from technical discussion forums," in *Proceedings of COLING*, Mumbai, India, December 2012.

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

[8] H. Hu, "Multimodal DBN for predicting high-quality answers in cQA portals," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, vol. 2, Sofia, Bulgaria, August 2013.

[9] 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.

[10] S. Scott and S. Matwin, "Text classification using WordNet hypernyms," in *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, Université de Montréal, Canada, August 1998.

[11] L. Feng, L. Wang, 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. 2018, pp. 493–498, 2018.

[12] E. Agichtein, "Finding high-quality content in social media," in *Proceedings of the 2008 International Conference on Web Search and Data Mining*, Palo Alto, CA, USA, February 2008.

[13] S. George K and S. Joseph, "Text classification by augmenting bag of words (BOW) representation with co-occurrence feature," *IOSR Journal of Computer Engineering*, vol. 16, no. 1, pp. 34–38, 2014.

[14] 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, no. 32, pp. 1–12, 2016.

[15] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," 2015, http://arxiv.org/abs/1509.01626.

[16] J. Huang, M. Zhou, and D. Yang, "Extracting chatbot knowledge from online discussion forums," in *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, Hyderabad, India, January 2007.

[17] S. D. 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] K. Chai, "Automatically measuring the quality of user generated content in forums," in *Proceedings of the Australasian Joint Conference on Artificial Intelligence*, Springer, Perth, Australia, December 2011.

[19] J. Jeon, "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*, Olympia, WA, USA, March 2006.

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

[21] D. Ö. Şahin and E. Kılıç, "Two new feature selection metrics for text classification," *Automatika*, vol. 60, no. 2, pp. 162–171, 2019.

[22] 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.

[23] V. Kharde and P. Sonawane, "Sentiment Analysis of Twitter Data: A Survey of Techniques," 2016, http://arxiv.org/abs/1601.06971.