MACHINE LEARNING METHOD FOR DETECTING ANSWERS IN DISCUSSION FORUMS: A BIG DATA ANALYTICS USE CASE

[1]Anushka Jain, [2]Satyam Khare, [3]Sanyam Bansal, [4]Saurav Chandra(Guide)
[1] KIET GROUP OF INSTITUTIONS,[2] KIET GROUP OF INSTITUTIONS, [3] KIET GROUP OF INSTITUTIONS, [4] KIET GROUP OF INSTITUTIONS
[1]Anushka.1923it1015@kiet.edu,[2] satyam.1923cs1065@kiet.edu, [3]sanyam.1923cs1028@kiet.edu, [4]saurav.chandra@kiet.edu

Abstract*— These days, data is pouring into internet discussion forums, and it would be ideal to transform the vast amounts of data into information that can be use. Online discussion boards are now an essential component of the internet and are important knowledge sources. This platform is used by users to ask questions and receive responses from other forum users. An introductory post (question) frequently receives multiple reply postings (answers), making it challenging for a user to quickly scan them all for the most pertinent and reliable response. Therefore, it is crucial to figure out how to autonomously gather the most pertinent response for a query within a thread. The task of response extraction is treated as a classification problem in this study.

Keywords: Online Discussion Forum, data analytics, machine learning.

I. INTRODUCTION

Online discussion forums have become increasingly popular in higher education as a tool for facilitating student engagement and promoting active learning (Boling et al., 2012). These forums provide a space for students to engage in discussions with their peers and instructors, share their perspectives, ask questions, and receive feedback.

In recent years, the MERN stack has emerged as a popular technology stack for building web applications, including online discussion forums. The MERN stack is a combination of four technologies: MongoDB, Express, React, and Node js.

We aim to develop an online forum "Talk It Geeks!" for group discussion. This is a web-based application for managing group discussion forums. Every time an end-user asks a question for information, the administrator receives it. Any user can post questions and answer other users' questions. There is a central database where all information is managed. Users can invite other users to discuss and make requests. Administrators have the authority to update the database. This is useful for small offices, schools, departments, or groups interested in effectively organizing. There are also affiliates of her users who act as intermediate users who, if they know, can answer questions of end-her users. Ability to share resources and publish articles for registered users to view. Updated by the end user as new information arrives.

The product is a brand-new self-contained website designed to act as a dialogue platform for KIET Ghaziabad students, allowing them to ask questions and discuss institute-specific and global issues. You can post questions, answer the questions, or request votes, users must create an account. This website's main goal is to give students a simple platform for conversation and interaction on a range of subjects. They can talk about academic and non-academic topics. They can even chat about other topics that are common to most people and offer their opinions on other atypical topics. The site can also help you seek help in finding blood donors among young people and keep records of the data collected.

II. LITERATURE REVIEW

A literature review of discussion forum websites emphasizes the importance of online communication and collaboration for student learning and engagement. According to research, university discussion forums can offer a welcoming and accepting atmosphere for students to express ideas, ask questions, and get feedback from instructors and peers. One study found that online discussion forums can encourage critical thinking and reflection by encouraging students to ponder their own opinions and respond constructively to others (Boling et al.., 2012).

Another study showed that discussion forums increased student engagement and satisfaction with courses by allowing students to interact with classmates and teachers outside the classroom (McCarthy & Higgs, 2019).

Additionally, literature suggests that university discussion forums can foster a sense of community and belonging among students, which is essential for academic success and well-being (Zhan & Mei, 2013). By providing a space for students to interact with peers and teachers, discussion forums can create a supportive and inclusive environment that encourages students to participate and contribute to the learning process.

Overall, the literature supports the use of discussion forum websites as a valuable tool for facilitating learning, engagement, and community building. Instructors can develop a more participatory and inclusive educational atmosphere that meets the variety of needs and perspectives of students by integrating discussion forums into their courses.

III. ADVANTAGES OF STUDENT DISCUSSION FORUM

Student discussion forums have several advantages:

- *Enhanced cooperation:* Student discussion forums provide a platform for students to collaborate and share ideas. It is an amazing platform to discuss course content, share resources, and feedback, and collaborate on group assignments, projects, and presentations.
- *Developing communication skills:* Students can hone communication abilities like communication through writing, analytical thinking, and troubleshooting by taking part in discussion forums. Students can learn how to express their thoughts clearly and effectively, receive feedback on their writing, and participate in discussions and discussions with peers.
- *Better learning:* Student discussion forums provide space for students to learn more about course content, ask questions, and receive feedback from peers and faculty. Students can learn from a variety of perspectives and experiences, expanding their knowledge and skills beyond those provided by lectures and textbooks.
- *Flexibility:* Online discussion forums offer flexibility as to when and where students can participate. Students can access the forums anytime, anywhere with an internet connection, so they can share information with their peers and complete assignments on their own schedule.
- *More engagement time:* Discussion forums can increase student engagement and motivation in your course

IV. ANALYSIS OF STUDENT DISCUSSION POSTS

Online discussion boards offer a considerably bigger pool of data for research and have the processing power to look at multiple variables simultaneously.

Quantitative analytics can be used in discussion forums to collect data on various aspects of the forum, such as user behavior, engagement, and content quality. Here's an example of how quantitative analysis can be used in a discussion forum.

- *Analysis of user behavior:* We collect data on the number of users, number of visits and time spent on forums. This information can be used to understand user engagement and identify areas for improvement.
- *Content analysis:* Collect data on the number of posts, comments, and votes (up-vote or down-vote) to understand the engagement and popularity of specific topics. This information may be used to identify the most popular topics and inform future content creation.
- *Sentiment analysis:* Collect sentiment data from posts and comments to understand sentiment across communities and identify areas of positive or negative sentiment.
- *User demographic analysis:* We collect data about user's demographic information, such as age of user, gender of the user, location, and degree of education.
- *Moderation analysis:* We collect data on the number of posts and comments moderated and the types of moderation actions taken to monitor the effectiveness of moderation and identify areas for improvement.

V. PROPOSED METHODOLOGY

Phase 1: Data is preprocessed in the first step to remove errors and noise.

Phase 2: The second phase calculates lexical and non-lexical attributes for the question and reply postings to ascertain similarity.

Phase 3: Several selection strategies are used to filter features.

Phase 4: In the last stage, the responses are categorized using the LinearSVC kernel method of SVM.

The following explains these steps.

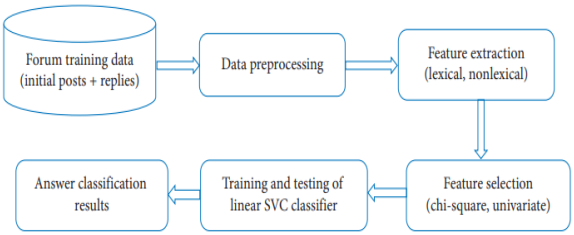

FIGURE 1: Proposed approach for answer detection in discussion forums.

1) *Preprocessing:* Data preparation entails transforming raw data into a format that can be predicted and analyzed. To preprocess the data, the following procedures are taken:
(a) Making every word lowercase
(b) Using WordNet to lemmatize wordsLemmatizer of NLTK
(c) Eliminating all stop words
(d) Extending the acronym

2) *Extraction of Features:* These characteristics, which can be grouped in many ways, are used to assess the relevance and similarity of a reply post to its original post. In their study, Osman et al. [1] categorized traits into six groups: politeness, creator activeness, accuracy, simplicity of comprehension, and amount of data. These categories were then divided into 28 sub characteristics.
The five feature groups—1) Content, 2) Lexical, 3) Forum-specific, 4) Reply-to types, 5) Structural were separated into 17 sub features in a different research study [6].
Lexical and non-lexical categories can be used to categorize features. Text-specific features include lexical features like the cosine matches of inquiry and reply posts. A similar lexical characteristic is the distinct word count of a reply post.
Examples of non-lexical features include forum-specific and content-based features. The number of topics a person has participated in overall, their position in the forum, and the interval between a question and a response post are examples of non-lexical factors.

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

For the purpose of mining answers from discussion forums, some studies advocate non-lexical traits above lexical ones, while others advocate lexical features.
Since questions and answers naturally contain similar words, it is advisable to use both lexical and non-lexical properties to identify the most pertinent and reliable response. While non-lexical features are used to assess an answer's standard, or the extent to which it relates to the question, lexical aspects are used to determine whether an answer is relevant to the question.
Certain functionalities are occasionally unavailable. One study looked at 12 data forums and discovered that 75% of author activeness features and 36.3% of forum-specific features are accessible [6]. Timeliness functions are not available in our situation. Additionally, using certain characteristics makes the model forum unique. In order to target features that are 100% available and simply calculable from the text or thread structure, we used both lexical and non-lexical features in this investigation. Lexical, content-driven, and semantic attributes make up these features.
In this investigation, twenty features indicated in Table-1 with an overview were used. Table-2 demonstrates that 14 of these characteristics are lexical, content-driven, or semantic attributes. Our new proposed semantic features are the three aspects in data table that are underlined: F1, F16, and F17. Some features, such as F20, F13, F12, F11, and F7, are based on text written or thread architecture. For instance, by breaking up a reply post into words and afterwards using the Len and Set functions in Python, it is possible to get F7, which is the count of unique phrases in a reply post.
We employed the bag-of-words (BoW) technique for pure lexical features like F6, F4, F5, F3, and F2. A well-known method for removing features from documents and representing them as vectors

is the BoW approach. The frequency of occurrence of a term in the documents is represented as a vector of values. We employed bigram and trigram word sequences for preserving the structure of sentences and order of words because the BoW approach disregards feature order and only considers term frequency, which will give the document additional depth of meaning.

To the best of our knowledge, no one has employed the three new semantic traits we introduced named F17, F1, and F16 for solution mining in discussion forums. We used Google's pre-trained word2vec model and word mover distance for our additional characteristics. The word2vec model will evaluate the importance of two phrases in a 300-dimensional space because we continue to use the default word vector length of 300 features.

Words with similar semantics or contexts will have near vectors, which is its specialty. Word mover distance (WM distance) is a metric used to compare two documents. The dissimilarity will be greater the closer the WM distance is, and vice versa. The Zero separation denotes a complete relationship between the two papers. Words with similar semantics or contexts will have near vectors, which is its specialty. The word mover (WM) distance calculates how different two documents are from one another. The dissimilarity will increase with increasing WM distance and vice versa.

According to Tables 3-6, both univariate and chi-square feature's selection procedures selected the three newly proposed semantic qualities (F17, F16, and F1) as the top most features for both the TripAdvisor (NYC) and Ubuntu datasets, making them the most important ones.

3*) Feature Selection:* To obtain answers from the question-answer forums, utilize a particular set of lexical and non-lexical properties. Not all of these, can be used for the following reasons:

(a) Certain characteristics are unimportant and have a negative impact on the performance of the model [1].

(b) Not every attribute can be found in datasets. Some features are obtained from feature combinations, while others are obtained from features that are correlated.

To get beyond the previously mentioned constraints, we first pick features that are always available and that are simple to calculate from the text. Then, in order to obtain the best features, we used feature selection techniques like univariate and Chi-square to reduce the feature space.

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

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

4) *Model Construction for Classification:* The goal of the Classification Model Construction phase is to use a machine learning algorithm to categorize the reply posts as significant relevant or irrelevant. LinearSVC, an SVM kernel approach, was utilized to classify the data. The categorization is based on how relevant a reply is to the original post.
We evaluated the LinearSVC classifier's classification accuracy to that of conventional SVM kernel methods and other cutting-edge classification algorithms like multinomial Nave Bayes, Bernoulli Nave Bayes, logistic regression and random forest, to name a few. Three sets of features, each consisting of two sub feature sets obtained using various feature selection strategies, were used to train and evaluate all classifiers. Section 4 has further information.

VI. EXPERIMENTAL SETTINGS

The suggested response detection algorithm is calculated using datasets, including the online Ubuntu Linux distribution forum and the TripAdvisor forum for New York City (NYC). The Ubuntu dataset has 756 total replies, and the TripAdvisor (NYC) dataset has 788 total replies. For answers, there are three categories that have been defined.
1)The class label 3 is given to responses that are entirely relevant.
2)The class label 2 to those that are just somewhat relevant
3)The class label 1 to those that are completely irrelevant.
We split the labelled dataset into training and testing portions, with 80% of the data used for each.
We employed support vector machine (SVM) techniques dubbed LinearSVC, which use a linear kernel, to categorize answer/reply postings in text forum threads. Text classification problems are typically resolved with SVM [22]. We quickly review the following classifiers. We also contrasted LinearSVC' s performance with that of other SVM kernel techniques and other state-of-the-art classification algorithms.

Naive Bayes: It is a set of learning algorithms(supervised) that treats each feature as a separate object and is based on the Bayes theorem. Many text classification problems have been successfully solved using this classifier. Bayes theorem is stated below:

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

where x_1 to x_n denotes a dependent feature vector and y is a class variable.
When compared to other classifiers, Naive Bayes trains quickly and takes little data.
The evaluation of this study makes use of the next iteration of Naive Bayes. Distributed multinomial data are used with Naive Multinomial Bayes. Its primary use is to categorize texts.

Support vector machines, or SVMs: It consists of a collection of algorithms for outlier recognition, regression and classification. In high-dimensional space, it works well and it needs less memory. It is compatible with a variety of kernels. Custom kernels can also be given. We applied the next three strategies.

- *The Support Vector Classification (SVC) method:* It is founded on libsvm. With more samples, the fit takes longer to complete. The standard kernel is "rbf". " Poly"," Linear " and "Sigmoid" are more kernel types.

- *NuSVC:* Although it has somewhat distinct mathematical techniques, formulations and parameter set from SVC, it is the same algorithm. It is based on libsvm. Regularization parameter Nu has values between 0 and 1. Although Nu evaluation is easier than C, C and Nu have equal classification power.

- *LinearSVC:* It is built using a kernel and "liblinear" library. The choice of penalties and loss functions is more variable when the input is sparse or dense.

Logistic regression: The application of logistic regression (LR) to problems involving several classes or over two discrete outcomes is extended using this classification technique. It is a model that projects the likelihood of potential results for a target variable using a collection of input dataset.

Radom Forests: Other names for them include random decision forests. For classification problems, they serve as an ensemble learning approach and build a lot of decision trees during the training period before generating the class which is related to the categorization of every one of the decision tree.

Feature Reduction: Two selection methods univariate and chi-square are used to remove redundant and unimportant features. The former chose eleven of the best attributes for Ubuntu and eight of the best for the TripAdvisor dataset, which are given in Tables 5 and 3, respectively, whereas the latter chose 15 of the best characteristics for Ubuntu and 10 of the best for the TripAdvisor dataset, which are shown in Tables 6 & 4, respectively.

Experimental Findings and Discussion: The results of all 6 classifiers used in the study using both all characteristics and features selected using various selection techniques are covered in this section. i) LinearSVC, ii) SVC, iii) NuSVC, iv) multinomial NB, v) random forest (RF), and v) logistic regression (LR) are the methods we use in this work.

In the initial phase, as indicated in Table 7, classifiers were applied to each of the twenty features from the Ubuntu dataset. The accuracy of 72.5% achieved by MultinomialNB and LinearSVC stood out among the six classifiers, which all gave good results. SVC's accuracy was 70.1%, while LR's accuracy was second-best (71.1%). The results of the evaluation of classifiers using the TripAdvisor dataset are shown in Table 8.

The second phase was condensing the feature space using the chi-square feature selection method. Tables 3 and 5 show the top 11 characteristics for Ubuntu and the 8 most prominent qualities for the TripAdvisor, respectively. The 3 new semantic characteristics that were introduced in this study were selected using the feature selection technique.

Table 7: Results for Ubuntu dataset using all features.

Classifiers	Accuracy
LinearSVC	72.5
SVC	70.1
MultinomialNB	**72.5**
Random forest	61.2
Logistic regression	71.1
NuSVC	60

Table 8: Results for TripAdvisor dataset using all features.

Classifiers	Accuracy (%)
LinearSVC	**67.2**
NuSVC	62.6
Random forest	65.1
Logistic regression	65.1
SVC	61.5
MultinomialNB	57

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

Classifiers	Accuracy (%)
LinearSVC	**73.4**
SVC	65.8
MultinomialNB	70.2
Logistic regression	70.2
Random forest	63
NuSVC	61.1

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

Classifier	Accuracy (%)
LinearSVC	**74.9**
NuSVC	66.7
Random forest	64.8
Logistic regression	71.6
SVC	63.7
MultinomialNB	61.8

Classifiers were applied to the Ubuntu dataset based on these desired features. Table 9 shows that LinearSVC, with a score of 73.4%, has the highest accuracy. MultinomialNB and LR are 70.2% accurate. SVC is ranked fourth with a 65.8% accuracy rating. Random Forest is in position five, and NuSVC is in position six. Regarding Tables 9 and 7, LinearSVC and LR offered accuracy that was on par the rest of the twenty-feature prototypes.

Together, Random Forest and NuSVC helped them become more accurate. Even though MultinomialNB' s accuracy was drastically reduced, only 11 as opposed to 20 characteristics were used this time. The top eight most effective features were selected using the chi-square approach, and Table 10 shows the results of the six classifiers LinearSVC, NuSVC, RF, LR, SVC, and MultinomialNB for the TripAdvisor dataset.

Again, LinearSVC performed well, outperforming LR and NuSVC with a 74.9% accuracy rate. The least accurate technology is RF (64.8%). LinearSVC accuracy rose by 7.7%, NuSVC accuracy rose by 4.1%, and LR accuracy rose by 6.5% when measured against the accuracy with all 20 characteristics mentioned in Table 8. Both SVC and MultinomialNB showed an increase in accuracy.

The third phase of the process involved filtering the features using the univariate selection of features method. The newly included three semantic traits were likewise present in the two datasets. Table 11 shows the classification outcomes for the Ubuntu dataset together with the selected traits. LinearSVC takes the lead with an accuracy of 75.2%. MultinomialNB' s precision is 72.5% respectively.

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

Classifiers
MultinomialNB
LinearSVC
SVC
Random forest
Logistic regression
NuSVC

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

Classifiers
LinearSVC
SVC
NuSVC
Random forest
Logistic regression
MultinomialNB

RF's accuracy is 60.1%, whereas SVC and LR classifiers have similar accuracy of 71.1%. The classifiers performed better with the chosen 15 features than with all 20 characteristics.
The top ten most effective features produced by the univariate feature selection method for the TripAdvisor dataset were used in algorithms. Table 12 shows that the classifiers perform much better than when using all twenty characteristics. While LR accuracy increased from 65.1% to 70.1%, LinearSVC accuracy increased from 67.2% to 72.9%. Figures 2 and 3 display, for the Ubuntu and TripAdvisor datasets, the classification accuracy of several classifiers based on various features. The results of the experiment led us to observe the following:
(1) The best picked features raised or maintained the accuracy of most of the classifiers.
(2) All other state-of-the-art classifiers were outperformed by our suggested classifier, LinearSVC.
(3) The two datasets selection approaches used our new three proposed semantic characteristics, which significantly increased the classification accuracy of LinearSVC.

Home page:

1. It serves as a gateway to the community and provides a welcoming and informative introduction of the forums' purpose and features.
2. Navigation bar consists of recent posts, user profile, settings, the search box.
3. There are different sections for the forum such as history, music, business, science, health, technology, etc.
4. The central part shows the feeds i.e., all the questions posted and the answers to them. There is an option to up-vote and down-vote the answers.

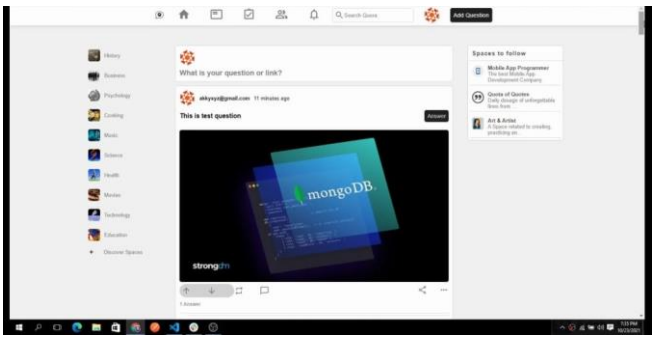

Figure 2

Post Question:
The user would be able to post any of his queries based on the forum.

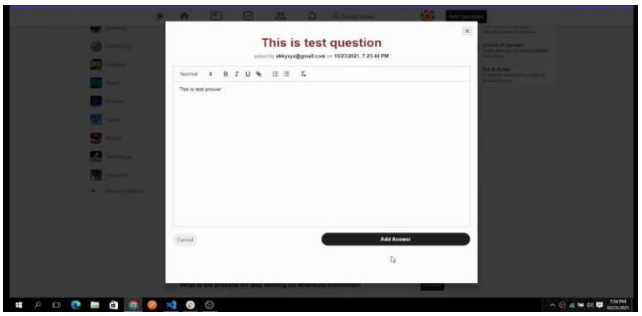

Figure 3

Answer Question:
Any user can answer any question. .

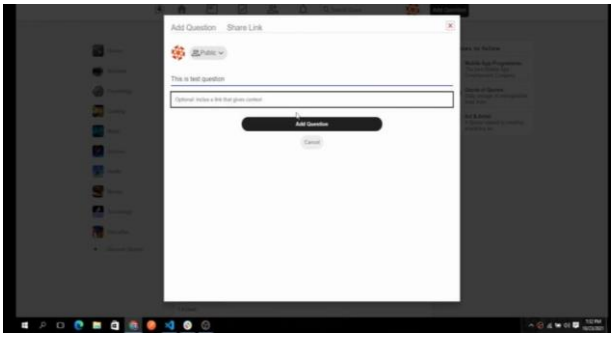

Figure 4

Logout:
The user can logout from the forum at any time the wish to.

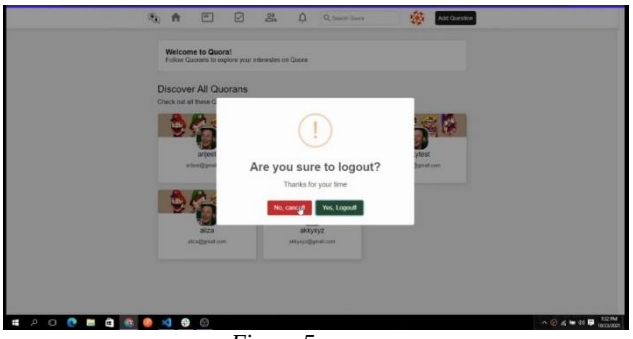

Figure 5

I. CONCLUSION

In a thread or discussion forum, it might be challenging to automatically select the most important and useful response to the first posting (question). This work adopts a novel method by outlining lexical, semantic, and content-based attributes that considerably improved the suggested classifier's classification accuracy. In this work, we suggested an SVM kernel method called LinearSVC for locating the most pertinent responses to the initial post in a forum thread, and we compared it to existing SVM kernel-based methods and other cutting-edge classification techniques. The most accurate version of SVM was LinearSVC. By examining two subsets of features, the model's performance was improved. Three more semantic characteristics were provided and chosen as the best features via univariate and chi-square approaches to decision-making, greatly boosting the accuracy of LinearSVC. In contrast to the univariate strategy, which chose 10 lexical and five non-lexical variables for the Ubuntu dataset, the chi-square method choose six lexical & five non-lexical qualities. The univariate methodology selected 7 lexical and 3 non-lexical features for TripAdvisor (NYC), while the chi-square technique chose 5 lexical and 3 non-lexical features. Thus, lexical characteristics turned out to be more significant and necessary for responding to questions on message boards.

To further enhance the model, we plan to investigate more semantic and content-driven components in the future.

REFERENCES

[1] Balakrishnan, G. & Coetzee, D. (2013). Predicting student retention in Massive Open Online Courses using Hidden Markov Models. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.659.2890&rep=rep1&type=pdf

[2] Beheshtiha, S. S., Hatala, M., Gašević, D., & Joksimović, S. (2016). The role of achievement goal orientations when studying effect of learning analytics visualizations. In Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (pp. 54-63). Edinburgh, United Kingdom: The University of Edinburgh.

[3] Boyer, S., & Veeramachaneni, K. (2015). Transfer learning for predictive models in Massive Open Online Courses. In C. Conati, N. Heffernan, A. Mitrovic, M. Verdejo (Eds.) Artificial Intelligence in Education. AIED 2015. Lecture Notes in Computer Science, vol. 9112. Springer, Cham.

[4] Crossley, S., Paquette, L., Dascalu, M., McNamara, D., & Baker, R. (2016). Combining click-stream data with NLP tools to better understand MOOC completion. In Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (pp. 6-14). Edinburgh, United Kingdom: The University of Edinburgh.

[5] Jiang, S., Williams, A. E., Schenke, K., Warschauer, M., & O'Dowd, D. K. (2014). Predicting MOOC performance with Week 1 Behavior. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.) Proceedings of the 7th International Conference on Educational Data Mining (pp. 273-275). London, United Kingdom: Institute of Education, UCL.

[6] Jimoyiannis, A., & Angelaina, S. (2012). Towards an analysis framework for investigating students' engagement and learning in educational blogs. Journal of Computer Assisted Learning, 28, 222–234.

[7] Science Direct /https://doi.org/10.1371/journal.pone.0225098

[8] Research Gate https://www.researchgate.net/publication/324417971_Analysis_of_student_discussion_posts_in_a_MOOC_Proof_of_concept

[9] Research Paper Link /https://doi.org/10.1016/j.compedu.2021.104402

[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, Universit´e de Montr´eal, 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-ofwords 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.