IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)
e-ISSN: 2278-2834,p- ISSN: 2278-8735.Volume 17, Issue 5, Ser. II (Sep. – Oct. 2022), PP 11-16
www.iosrjournals.org

Sentiment Analysis Using Machine Learning

Lingaraj D M[1],Mohammed Adnan Hakeem[2], Mohan Krishna B[3], Mohith Kumar[4],Deepika D Pai[5]

[1](Student,Electronics and Communication Engineering, Vemana Institute of Technology/ VTU, India)
[2](Student,Electronics and Communication Engineering, Vemana Institute of Technology/ VTU, India))
[3](Student, Electronics and Communication Engineering, Vemana Institute of Technology/ VTU, India)
[4](Student, Electronics and Communication Engineering, Vemana Institute of Technology/ VTU, India)
[5](Assistant Professor, Electronics and Communication Engineering, Vemana Institute of Technology/ VTU, India)

Abstract:
Educational Data Mining is a prominent area toexplore information in educational fields using data miningalgorithms. In this paper, we have used a few learning algorithmsto effectively rate the faculty belonging to an educational instituteon the basis of feedback submitted by the students. Our proposedmodel uses sentimental analysis and machine learning classifieralgorithms for capturing the emotions from the student'sfeedback system. This model gives an accurate and efficient wayto rate the faculty belonging of a particular educational institute.Sentimental Analysis is a reference to the task of natural languageprocessing to determine whether a text contains subjective information and what information it expresses i.e.,whether the attitude behind the text is positive, negative orneutral, examine the sentiments present in the text document forclassification of students' feedback based on polarity (positive/negative/ neutral) using machine learning and NLP methods.

Key Word**: **Educational Data Mining,sentimental analysis, classifier algorithms,NLP methods

---

Date of Submission: 03-10-2022 Date of Acceptance: 17-10-2022

I. Introduction

Sentiment analysis is a process where the dataset consists ofemotions, attitudes or assessment which takes into account the way ahuman thinks. In a sentence, trying to understand the positive and thenegative aspect is a very difficult task. The features used to classify thesentences should have a very strong adjective in order to summarizethe review. These contents are even written in different approacheswhich are not easily deduced by the users or the firms making itdifficult to classify them.

Sentiment analysis influences users to classify whether the information about the product is satisfactory or not before theyacquire it. Marketers and firms use this analysis to understand their products or services in such a way that they can be offered as per theuser's needs. There are two types of machine learning techniqueswhich are generally used for sentiment analysis,one is unsupervisedand the other is supervised Unsupervised learning does not consist ofa category and they do not provide the correct targets at all and therefore conduct clustering.Supervised learning is based on a labelled dataset and thus the labels are provided to the model during theprocess. These labelled datasets are trained to produce reasonableoutputs when encountered during decision-making.

II. Literature Survey

The literature review provides inspiration for developing a module.There have been various studies conducted on the topic of applicationselection. In the first part, we looked at some early research that wasdone to find the optimum methodology. The different protocolsemploying diverse methodologies were investigated and explainedafter analyzing papers on related issues.

The literature review serves as a springboard for creating a module.Several studies on the topic of applicant selection have been undertaken.We looked at some early research that was done to discover the bestmethodology in the first half. After reviewing publications on connectedtopics, the many protocols employing various approaches were studiedand discussed.

When Educational Data Mining is a prominent area to exploreinformation in educational fields using data mining algorithms. Fewlearning algorithms have been used to effectively rate the facultybelonging to an educational institute on the basis of feedback submitted by the students. The proposed model uses sentimental

analysis andmachine learning classifier algorithms for capturing the emotions fromthe student's feedback system. This model gives an accurate andefficient way to rate the faculty belonging of a particular educational institute. With this proposed model the faculty will be evaluated andrated with certain specified parameters which will help us to improvethe academic and education standards.

Several machine learning techniques which are used in analyzing sentiments and in opinion mining. Sentimental analysis with the blendof machine learning could be useful in predicting product reviewsand consumer attitudes towards newly launched products. The studyshows a detailed survey of various machine-learning techniquesand thencompared their accuracy, advantages and limitations ofeach technique. On comparing we get 85% of accuracy by using the supervised machine learning technique which is higher than that ofunsupervised learning techniques. Another study presents a combination of machine learning andlexicon-based approaches for sentiment analysis of students' feedback.The textual feedback, typically collected towards the end of asemester, provides useful insights into the overall teaching quality andsuggests valuable ways for improving teaching methodology. Asentiment analysis model is trained using TF-IDF and lexicon-basedfeatures to analyze sentiments expressed by students in their textualfeedback. A comparative analysis is also conducted between theproposed model and other methods of sentiment analysis. Theexperimental results suggest that the proposed model performs betterthan other methods.

III. System Overview

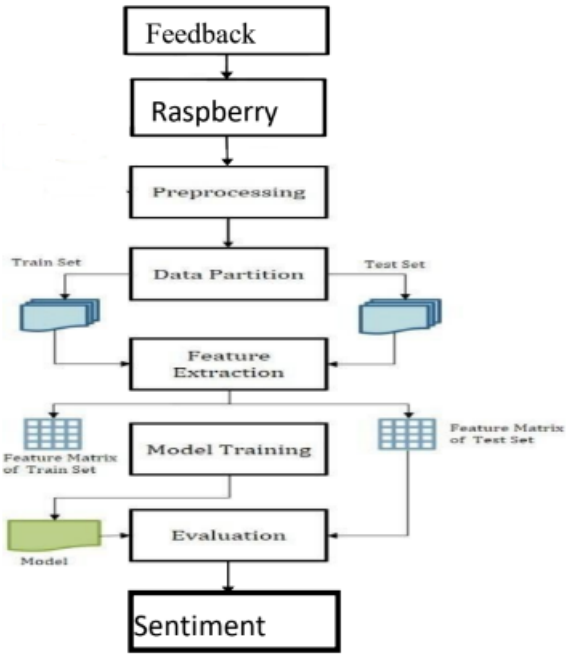


Fig 1: Block diagram of sentiment analysis

IV. Design Methodology

The presented methodology classifies sentiment polarity as positive, negative and neutral. The process workflow is shown in Fig. 1 and is further described in the following subsections.

A. Dataset Description

The dataset used in this paper comprisesof a number of comments extracted from various portals. The dataset is manually labelled with sentiment polarity labels {positive, negative, neutral}.

B. Preprocessing Student feedback data represents an unstructured text. To extract useful information from theunstructured text, several preprocessing steps should be applied to remove spelling errors, grammatical mistakes, URLs, etc. from the text. During the preprocessing stage, the following tasks should be performed using Python's NLTK library for preprocessing.

1) Punctuations: Punctuations, numbers and other special characters should be removed as these characters do not carry useful information related to sentiment analysis.

2) Tokenization: Tokenization is the process of splitting a text stream into a list of words.

3) Case Conversion: After tokenization, words are transformed into lowercase.
4) Stop words: In natural language processing, words that are frequently used such as helping verbs, prepositions, and articles are termed stop-words. Stop-words generally do not provide any useful information and therefore were removed from the feedback text.

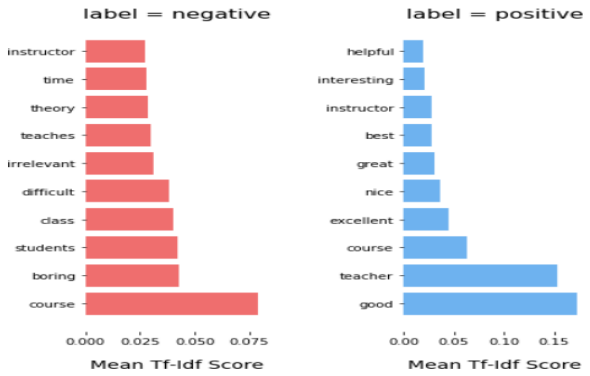
C. Data Partition For training and evaluation purposes, the manually labelled dataset will be randomly split into a train set and a test set. 70% of the dataset will be used for training and the remaining dataset used for evaluation purposes. Table I presents the distribution of sentiment labels in the training and testing datasets.

DISTRIBUTION OF SENTIMENT LABEL IN DATASETS

Sentiment Label	Trainset (sentences)	Testset (sentences)
positive	713	306
negative	126	55
neutral	22	8

Table 1:Distribution of sentiment labels

D. Feature Extraction After data splitting, feature extraction will be applied on both training and testing datasets. During the feature extraction stage, the preprocessed text is converted into a numerical feature vector using Ngram and TF-IDF.
Term Frequency-Inverse Document Frequency (TF-IDF): TF-IDF metric determines the importance of a word to a document in a given corpus. It assigns a higher weight to the words that occur frequently in a set of documents labelled with a particular sentiment polarity but least occur in a corpus.



Top 10 words in positive and negative comments using TF-IDF metric

E. Training Model
After the extraction of features from the train and test dataset,learning algorithms were applied tothe training model. The hybridmodel for sentiment analysis was trained using TF- IDFfeatures.
A brief description of the learning algorithms is given below:
Random Forest: Random Forest Algorithm was proposed in thisstudy, and scikit-learn implementation of the Random Forest algorithmwas used. The hyperparameters were tuned using three-fold cross-validation.
Support Vector Machines (SVM): The scikit-learnimplementation of SVM with a linear kernel was used to trainthe model.
Naïve Bayes Classifier (NBC): The basic idea involved in the naïve Bayes classification technique is to find the classes probabilitiesassigned to texts by using joint probabilities of classes andwords. The features/predictors used by the classifier are thefrequency of the words present in the dataset.

V. Results and Discussions

A. Processing of Feedback:

Stop words are removed and words are formed into lists.

```
custom_FB = "The Teacher is good but, her teaching method is not good. Students wont understand her methods ! "

print(process_feedback(custom_FB))
```

```
['teacher', 'good', 'teach', 'method', 'good', 'student', 'wont', 'understand', 'method']
```

B. Accuracy of Training models

i. Logistic Regression Method

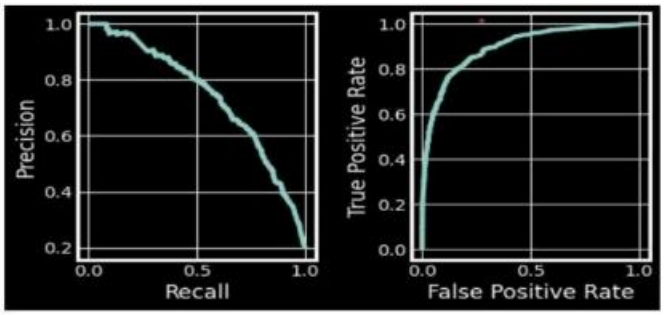
```
Train confusion matrix is:
[[6829   26]
 [   5 1795]]

Test confusion matrix is:
[[2215  108]
 [ 238  325]]

Train accuracy score:  0.996418255343732
Test accuracy score:  0.8801108801108801

Train ROC-AUC score:  0.9982442661479861
Test ROC-AUC score:  0.8956867344777572

Are under Precision-Recall curve: 0.6526104417670683
Area under ROC-AUC: 0.7441899264879837
```



ii. Naïve Bayes Method

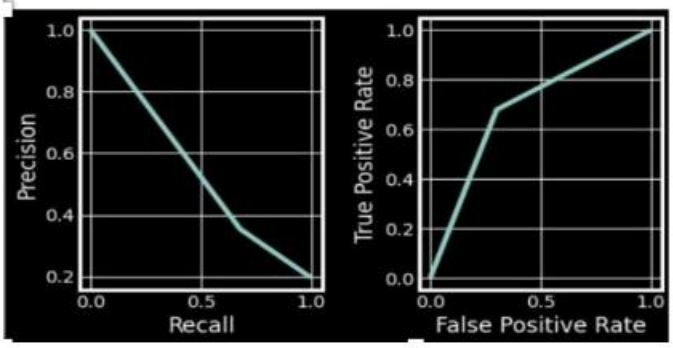
```
Train confusion matrix is:
[[5543 1312]
 [   0 1800]]

Test confusion matrix is:
[[1623  700]
 [ 181  382]]

Train accuracy score:  0.8484113229347198
Test accuracy score:  0.6947331947331947

Train ROC-AUC score:  0.9043034281546316
Test ROC-AUC score:  0.688586755810495

Are under Precision-Recall curve: 0.4644376899696049
Area under ROC-AUC: 0.5471372315951626
```



iii. NLTK Method

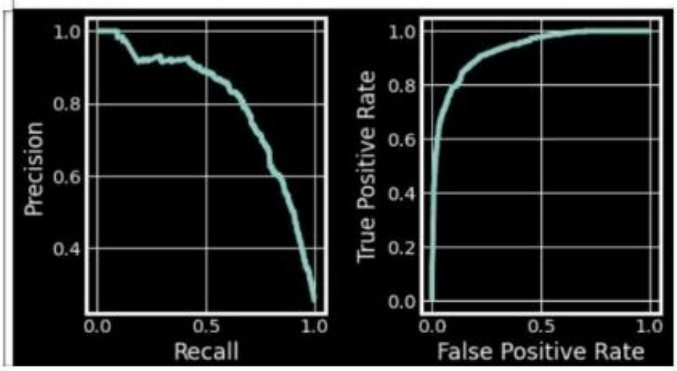
```
Train confusion matrix is:
[[6824   31]
 [ 151 1649]]

Test confusion matrix is:
[[2291   32]
 [ 296  267]]

Train accuracy score:  0.9789716926632005
Test accuracy score:  0.8863478863478863

Train ROC-AUC score:  0.9969059080962801
Test ROC-AUC score:  0.9291753099937378

Are under Precision-Recall curve: 0.6194895591647333
Area under ROC-AUC: 0.8049838180297202
```



C. Outputs of Different Methods

```
Feedback = input("Enter your review : ")
print(process_tweet(Feedback))
y_hat = predict_tweet(Feedback, freqs, theta)
print(y_hat)
if y_hat > 0.5:
    print('Positive sentiment')
else:
    print('Negative sentiment')
```

```
Enter your review : Punctuality is good. Most of the teachers are not interacting with students .
['punctual', 'good', 'teacher', 'interact', 'student']
[[0.5134384]]
Positive sentiment
```

ii. Naïve Bayes Method

```
feedback = ' Teachers arent giving depth of course.They trying to cover there work only'

p = naive_bayes_predict(feedback, logprior, loglikelihood)
print(p)

if p>0:
    print("Postive Feedback")
else :
    print("Negative Feedback ")
```

```
-0.41080248579296746
Negative Feedback
```

iii. SVM Method

```
test_sentence1 = "Punctuality is good,most of the teachers are  interacting with students"
predict_sentiment(test_sentence1)
```

```
Predicted label:  positive
```

VI. Conclusion

Sentimental analysis has become a popular research area due to the increasing number of internet users, social media etc. This extracted new features that have a strong impact on finding the polarity of the movie reviews. Then to perform the feature impact analysis by estimating the information gained for each feature in the feature set and using it to derive a reduced feature set. The main goal of this work is to classify the sentences according to their sentiment by using the Random Forest classification technique. This process of extracting the text having sentiment deals with finding the sentiment feature set from the sentences. As the final output is displayed it becomes easier for the user to understand the exact polarity result. In future work apply the concept of NLP in more detail for better prediction of the polarity results. To use the best classification technique for achieving the highest accuracy. This technique can also be implemented in other domains of opinion mining such as product reviews, political discussion forums, hotels, tourism etc.

References

[1]. L.Kousalya and R.Subhashini, "Sentimental Analysis for Students' Feedback using Machine Learning Approach", International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2395-0056, Volume-06 Issue-04 -04-2019 .

[2]. Daneena Deeksha Dsouza, Deepika, Divya P Nayak," Sentimental Analysis of Student Feedback using Machine Learning Techniques ", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1S4, June 2019.

[3]. K. S.Krishnaveni, Rohit R Pai, V. Iyer, "Faculty Rating System Based on Student Feedbacks Using Sentimental Analysis", International Conference on Advances on Computing, Communication and Informatics, pp.1648-1653, 2017.

[4]. B. K.Bhavitha, A. P. Rodrigues, N. N Chiplunkar, "Comparative Study of Machine Learning Techniques in Sentimental Analysis". International Conference on Inventive Communication and Computational Technologies, pp.216- 221, 2017.

[5]. Zarmeen Nasim, Quratulain Rajput and Sajjad Haide," Sentiment Analysis of Student Feedback Using Machine Learning and Lexicon Based Approaches", Institute of Electrical and Electronics Engineers (IEEE),e-ISSN:2324-8157

Lingaraj D M, et. al. "Sentiment Analysis Using Machine Learning." *IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)* 17(5), (2022): pp 11-16.