Pakistan Journal of Engineering Technology and Science (PJETS)
Volume 11. Issue. 2, PP. 32-40, December 2023
https://doi.org/10.22555/pjets.v11i2.1039
E-ISSN 2224-2333
P-ISSN 2222-9930

STUDENTS TEXTUAL FEEDBACK SENTIMENT ANALYSIS USING MACHINE LEARNING TECHNIQUES TO IMPROVE THE QUALITY OF EDUCATION

Naseer Ahmed[1], Mansoor Ahmed Khuhr[2], *Afrasiyab Khan[3], Muhammad Dawood[1], Mazhar Ali Dootio[4], Najeeb Ullah Jan[5]

[1]Department of Computer Science, Balochistan University of Engineering & Technology, Khuzdar, Pakistan
[2]Department of Artificial Intelligence and Mathematical Sciences, Sindh Madresatul Islam University Karachi, Pakistan
[3]Department of Software Engineering, Balochistan University of Engineering & Technology, Khuzdar, Pakistan
[4]Department of Computer Science and Information Technology, Benazir Bhutto Shaheed University Lyari, Karachi, Pakistan
[5]Department of Computer System Engineering, Balochistan University of Engineering & Technology, Khuzdar, Pakistan

[*]Corresponding author: *afrasiiyabkhan@gmail.com*

Abstract: **Several educational institutions worldwide work hard to obtain student feedback to explore their views on the courses and faculty. This feedback is utilized to enhance the institution's environment. In this modern world, institutes use data or feedback collection techniques. Still, they do not have proper techniques to analyze and use this data to improve the institute's educational quality from such textual feedback. This study presents techniques for analyzing the sentiments of student textual feedback. In this paper, machine learning methods, including Random Forest, Multinomial Nave Bayes Classifier, and Long Short-Term Memory, are applied. These methods are compared, and the experimental findings show that Long Short-Term Memory provides higher accuracy which is 95.75%. This model successfully predicts the sentiments of three types with high accuracy 1) Positive 2) Neutral 3) Negative.**

Keywords: **Textual feedback; Sentiment analysis; Machine Learning**

I. INTRODUCTION

In recent days, Sentiment analysis has become more popular as more companies pay close attention to reviews[1]. One of the critical NLP techniques is sentiment analysis [2]. Sentiment analysis is indispensable when evaluating people's emotions and providing polarity results. Reviews on various subjects, such as products and personalities, hold great significance in the eyes of organizations, which is why appropriate text structuring proves helpful for understanding the sentiment status[3]. Feedback provided by students can be categorized into two forms: textual feedback and grading feedback based on Likert scale scores. In the case of Likert scale scoring, students are presented with questions and asked to rate their responses using a predefined scale. This approach primarily concentrates on gathering feedback related to specific topics, but it may not accurately capture the perfect/present sentiments of the students[4]. Textual feedback is used to determine the actual sentiment of the students. Students are given a series of questions to reply in sentences in this textual style. It benefits both the academic administration and the instructor in

Pakistan Journal of Engineering Technology and Science (PJETS)
Volume 11. Issue. 2, PP. 32-40, December 2023
https://doi.org/10.22555/pjets.v11i2.1039
E-ISSN 2224-2333
P-ISSN 2222-9930

overcoming organizational challenges. Google Forms are used in this work to collect student responses with varying opinions. The objective is to extract statements of opinion and use machine learning techniques to identify them as either positive or negative. Machine learning approaches may use supervised or unsupervised learning. Classification issues may be resolved using several techniques, including Long short-term memory, naive Bayes, and random forests. The lexicon-based method detects sentiment polarity in textual information by utilizing a sentiment lexicon, essentially a collection of terms with associated sentiment polarities. This Paper is organized as follows: We summarize previous research on sentiment analysis and machine learning methods in the "Literature Review" section. The "Methodology" section describes categorizing material using student comments. The "Performance Analysis" section contrasts machine learning techniques, F-score, and accuracy. The "Conclusion" section summarizes the findings and offers final observations.

II. LITERATURE REVIEW

To improve the quality of education, many sectors around the world are currently focusing on this issue, making it one of the most trending topics nowadays. This improvement intrinsically depends on teaching efficiency. Thus, accurate and efficient evaluation of teaching has become an important area of academic research [5]. The area of sentiment analysis has been extensively researched. The text classification field hasn't seen many studies that classify texts into negative, positive, and neutral categories [6]. There are three types of methods used in sentiment analysis; (1) lexicon-based approaches, where a sentiment dictionary is created to evaluate the sentiment of words, (2) machine learning approaches that utilize manually designed features to train a non-neural network classifier, for categorizing words based on their sentiment and (3) deep learning approaches that involve training an advanced neural network model to capture more meaningful and abstract semantic features, for sentiment analysis[7]. Method three will be used in this paper. A thorough survey was conducted to evaluate sentiment in three important areas, including framework, feature extraction, and sentiment analysis [8]. As typical approaches, supervised learning methods such as Naive Bayes and Long short-term memory (LSTM) are examined. The results show that LSTM outperforms other classifiers in terms of accuracy. According to the study, Naive Bayes outperforms LSTM when dealing with small datasets, whereas LSTM outperforms when dealing with large datasets. [9]. A data mining methodology is developed to classify an institution's faculties ranking from 1 to 5 based on particular features. The Paper used the Naive Bayes classifier and text-mining algorithms to analyze student comments. However, one disadvantage of this study was that it did not accurately reflect the student's actual feelings [10]. In 2014, a sentiment analysis classification model was created specifically for Arabic text. After pre-processing, 2591 texts out of 10,500 were selected for training the model. The Naïve Bayes, SVM, and KNN classifiers were employed using the 10-fold cross-validation method to assess the sentiment of the reviews. The SVM classifier demonstrated the highest accuracy, reaching 75.25% [11]. In 2023, the automatic scoring model used for teaching evaluation also provided low accuracy results, which are discussed in a paper that is almost 79% accurate [5]. In the same year, another researcher achieved an accuracy of 63.70%, which is not bad but may not yield satisfactory results [12]. It can take time and effort to process a significant amount of feedback gathered towards the end of the semester. These are the major reasons that help us identify a research gap and develop a more accurate textual analysis model.As described in Long Short-Term Memory (LSTM), Naive Bayes, and Maximum Entropy (ME), the machine learning techniques are [9] [10]. Multinomial Naive Bayes and decision trees are used to analyze sentiment in Twitter data [9]. The n-gram approach was used to extract features from 1150 documents used in the study. A comprehensive research conducted by (Jin Zhou and Jun min Ye) has allowed us to embark on something. They diligently explored five databases and discovered 41 relevant articles. The findings indicate a focus, on education in most studies with an inclination, towards the utilization of smaller datasets. This motivated us to take a step and gather our dataset [13] . Education is a major area because a wide range of research

Pakistan Journal of Engineering Technology and Science (PJETS)
Volume 11. Issue. 2, PP. 32-40, December 2023
https://doi.org/10.22555/pjets.v11i2.1039
E-ISSN 2224-2333
P-ISSN 2222-9930

gaps is found in it almost most of the sectors adopted this feature and models are also developed by using Chinese stock reviews [14]. The classification model's performance was assessed using recall, F-measure, precision, and accuracy measures [15]. The problem of sentiment polarity categorization is addressed as input data; the study leverages online product reviews from Amazon [16]. During the COVID-19 pandemic, researchers developed a sentiment analysis model by analyzing tweets from several social media users. This model, which incorporates LSTM technology has proven to be quite effective, with an accuracy rate of 93% [17] In parallel another model was created using 17,155 tweets related to e-learning to gain insights and enhance it within the context of the pandemic[18]. Some experts have raised concerns, about the text feature extraction capabilities of LSTM-based methods. As an approach, they proposed a task aspect category sentiment analysis model based on RoBERTa (Robustly Optimized BERT Pre-training Approach). Although this method shows promise satisfactory results have not yet been achieved. Hence our focus remains on improving the conventional LSTM method [19].

III. METHODOLOGY

QEC of Balochistan University of Engineering and Technology Khuzdar collected student feedback at the end of every semester. We get help from the QEC department which provides us with Survey comments that are used to train the algorithm. When provided with test samples, this data trains the system using machine learning methods, allowing it to categorize texts into three categories Negative, positive, and neutral. The classification findings are visualized using a graphical form. This method includes six critical processes, as illustrated in Figure 1:

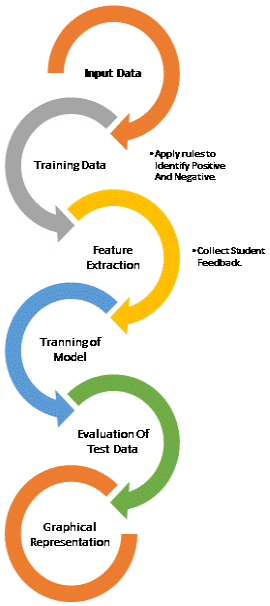

Figure 1: Sentiment Analysis Working

Gathering student feedback, preparing training data, extracting pertinent features, training the model, going through test data, and graphically presenting the results. These results help to use student textual feedback to improve the quality of education in all sectors.

Data from QEC surveys from Balochistan UET Khuzdar is used to carry out these steps.

A.Dataset

A collection of textural feedback from the students of BUET Khuzdar University from 2019 to 2022 served as the dataset for this study. The feedback was collected from the QEC (Quality Enhancement Cell). The feedback text was pre-processed to remove stop words and punctuation. The feedback's sentiment was then divided into three groups: positive, negative, and neutral. The dataset was divided into a training set and a test set. The training set was used to

Pakistan Journal of Engineering Technology and Science (PJETS)
Volume 11. Issue. 2, PP. 32-40, December 2023
https://doi.org/10.22555/pjets.v11i2.1039
E-ISSN 2224-2333
P-ISSN 2222-9930

train the LSTM machine learning model. Using the test set, the model's effectiveness was evaluated. The dataset was limited in size and diversity. We also used an open-source dataset from https://github.com/Afrasiyab-khan/Sentiment-Analysis-Research-Improve-the-quality-of-Education-by-Feedbacks.git to improve the model's performance. The Kaggle dataset included student feedback from a variety of universities around the world.

B. PreparingTtraining Data

The two types of machine learning are supervised learning and unsupervised learning. In supervised learning phrases or data points are labeled with a class, while unsupervised learning does not use labeled data. The system is trained using gathered training data, with each question having its own training data set. Students' sentence replies provide the training data. Sentiment Intensity Analyzer () from the Vader Sentiment package is used to analyze the sentiment of each phrase. Based on the "Valence Aware Dictionary and Sentiment Reasoner" (VADER), this technique provides a sentiment score between -1 and 1, signifying negativity to positive. Each word in the phrase is awarded a sentiment score ranging from -4 to 4 [4].

C. *Student Feedback Collection*

The data or feedback are collected from the "Balochistan University of Engineering and Technology Khuzdar" students in the form of questions that QEC collects at the end of every year using the University website, shown in Figure 2. The feedback QEC was collected in the form of a Microsoft Xcel sheet.

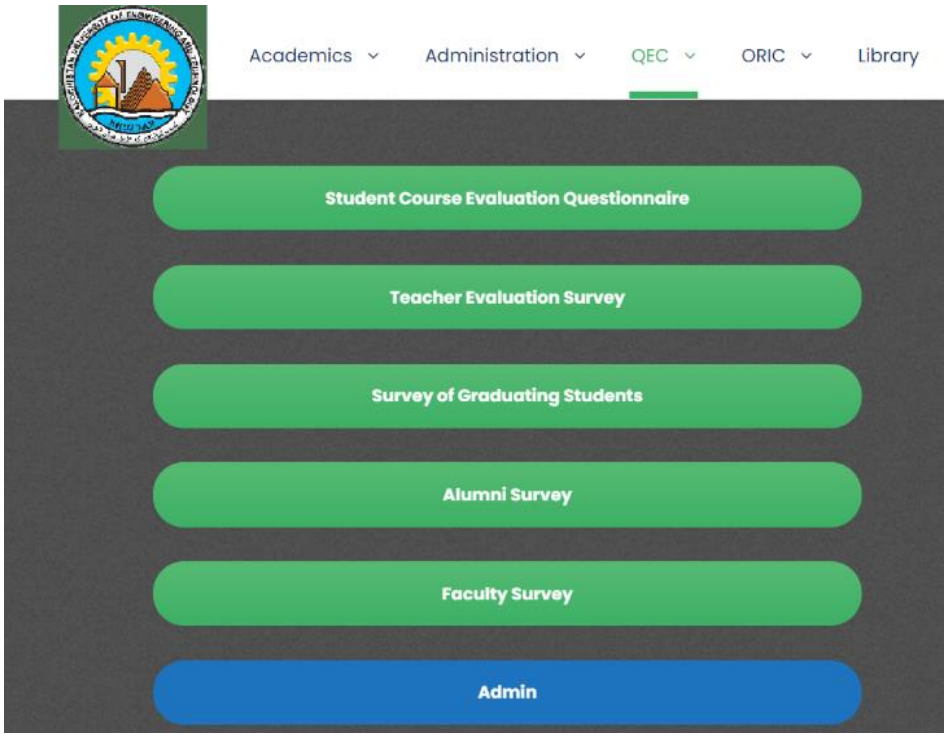

Figure 2: Data Collected From QEC Department BUET

D. Feature Extraction

Feature extraction is conducted on the datasets in this procedure to generate a format appropriate for machine learning algorithms. Feature extraction is used for both training and test sets of data. Tools for tokenizing textual data and extracting features are both included in the Scikit-learn package. Textual data is tokenized, separating it into individual words or tokens. The scikit-learn Count Vectorizer tool is used to tokenize text documents and generate a vocabulary of recognized terms.

Pakistan Journal of Engineering Technology and Science (PJETS)
Volume 11. Issue. 2, PP. 32-40, December 2023
https://doi.org/10.22555/pjets.v11i2.1039
E-ISSN 2224-2333
P-ISSN 2222-9930

E. Model Training

This study focused on long short-term memory, multinominal naive Bayes, and random forest approaches for text classification using machine learning.

Long Short-Term Memory (LSTM): Machine translation tasks frequently use Long Short-Term Memory (LSTM), a recurrent neural network (RNN). LSTMs are designed to handle sequential data, such as text, and can maintain a memory of previous inputs over a more extended period than traditional RNNs[20].

Multinomial Naïve Bayes: The basic objective of the naive Bayes classification method is to categorize texts using a combination of class and word probabilities. The frequency of the terms in the document is one of the characteristics or predictions used by the classifier [21][22] [23].

Random Forest: Random forest (RF) is a reliable machine-learning classifier that can be applied to both classification and regression applications. Its primary advantages are its non-parametric character, outstanding classification accuracy, and capability to comprehend variable importance[24].

Evaluation of the test data: The collection and analysis of data to evaluate an organization's success in carrying out planned operations is what evaluation entails. In the context of a model, assessment refers to the last step after training is completed. This stage is critical for determining the model's performance and generalizability. A separate test set is used to examine the trained model's accuracy and performance during the assessment. By applying it to the test set, we may assess the model's capacity to predict outcomes on fresh, previously unknown data reliably. This assessment procedure offers information on the model's working correctness and ability to make predictions beyond the data it was trained on. The assessment phase is crucial in establishing the model's dependability and potential for practical implementation. It aids in validating the model's performance, identifying development areas, and determining its acceptability for real-world use.

Performance Analysis: This section discusses Multinomial Naive Bayes, Random Forest, and Long Short-Term Memory, three machine learning methods. Based on the accuracy and F-score measures, the analysis evaluates the effectiveness of these algorithms using unigram and bigram features. Unigrams are single elements or tokens taken from a string, whereas bigrams are sequences of two elements or tokens from the same string. Bigrams and unigrams can both be used as characteristics to evaluate and compare the performance of machine learning systems. Using a test dataset, the learned model's performance is evaluated. Multiple factors are taken into account when evaluating its effectiveness. Accuracy and F Score, two important metrics, are used to quantify the performance of the model. The level of accuracy indicates how accurate the model's predictions are on the whole. Precision and recall are balanced by a statistic called the F Score, also called the F1 Score. It weighs the trade-off between false positives and false negatives as well as the proportion of accurate forecasts. When the class distribution within the dataset is unbalanced. Accuracy: Equation 1 states that accuracy is calculated by dividing the total number of rows in the dataset by the number of accurate predictions the model produced[25].

$$Accuracy = \frac{\text{characters/words correctly recognized}}{\text{All characters/words}} \quad (1)$$

Calculating the F-score: an integral measure when evaluating models- requires accounting for both recall and precision together. This vital metric utilizes equation 2 as its formula and considers a weighted average approach that enables an accurate determination of how well models perform[26]

$$F\ Score = f\ 1 = f = 2 * \frac{precision \times Recall}{Precision + Recall} \quad (2)$$

Pakistan Journal of Engineering Technology and Science (PJETS)
Volume 11. Issue. 2, PP. 32-40, December 2023
https://doi.org/10.22555/pjets.v11i2.1039
E-ISSN 2224-2333
P-ISSN 2222-9930

Figures 3,4 and Table 1 display the accuracy results for both unigrams and bigrams using the Random Forest, Long Short-Term Memory, and Multinomial Nave Bayes Classifier. The training set size was varied while the test data was kept constant for the research. Notably, the models' accuracy increases as the training data increases.

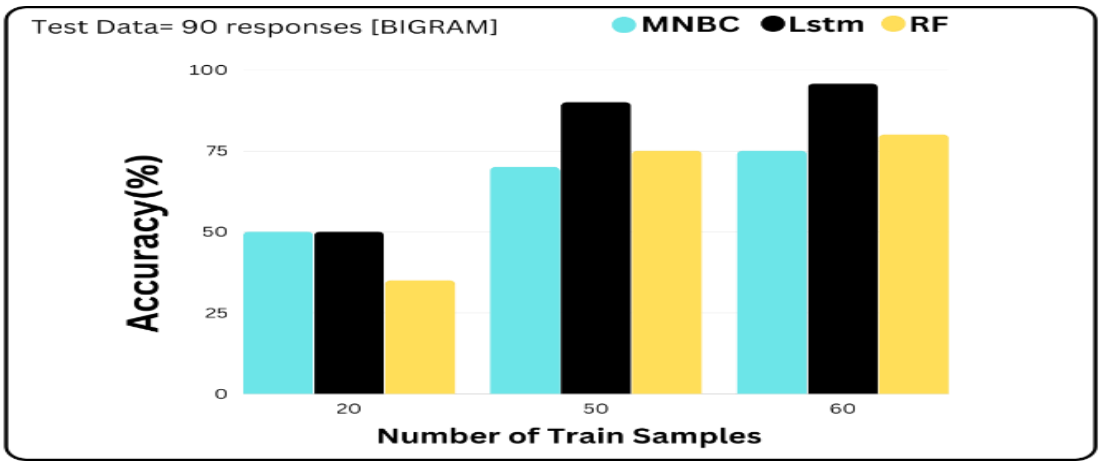

Figure 3: Performance of the Bigram LSTM, MNBC, and RF algorithms for various train samples

Table 1. Accuracy of diffrent methods

Model	**Accuracy**	**Precision**	**Recall**	**F-Score**
LSTM	95.75%	92.5%	90%	85.4%
MNBC	75.6%	82.5%	70.5%	70.4%
RF	80.2%	80.5%	38.6%	60%

When more data is used to train the model, the LSTM approach regularly provides higher accuracy. As compared to the Multinomial Naive Bayes Classifier (MNBC), Random Forest (RF), and other algorithms in terms of accuracy. On the other hand, RF and MNBC's accuracy is not improved. The LSTM provides greater accuracy as a result than

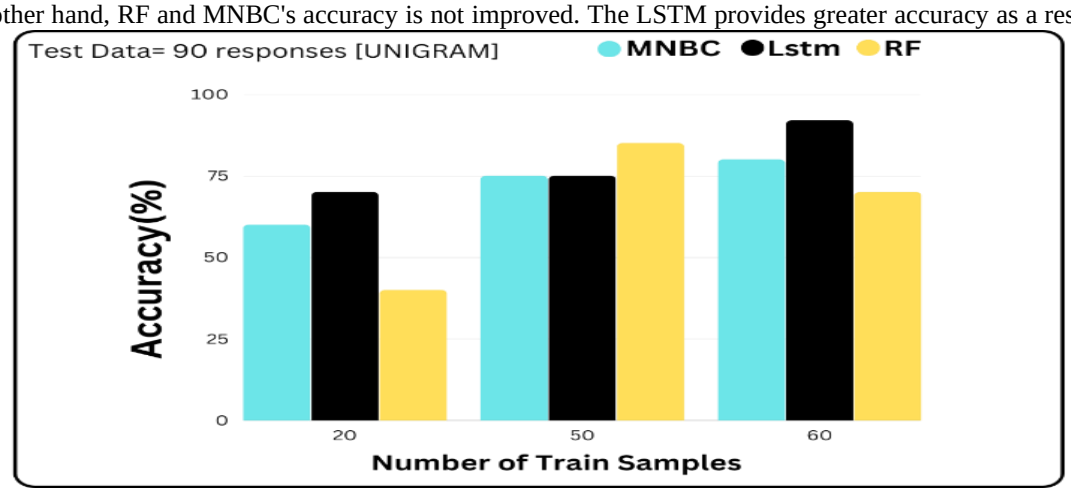

Figure 4: Displays the performance of the LSTM, MNBC, and RF algorithms using different train samples.

MNBC and RF.

Pakistan Journal of Engineering Technology and Science (PJETS)
Volume 11. Issue. 2, PP. 32-40, December 2023
https://doi.org/10.22555/pjets.v11i2.1039
E-ISSN 2224-2333
P-ISSN 2222-9930

The MNBC and LSTM algorithms' F Scores rise linearly with the amount of training data shown in Figures 5 and 6 respectively.

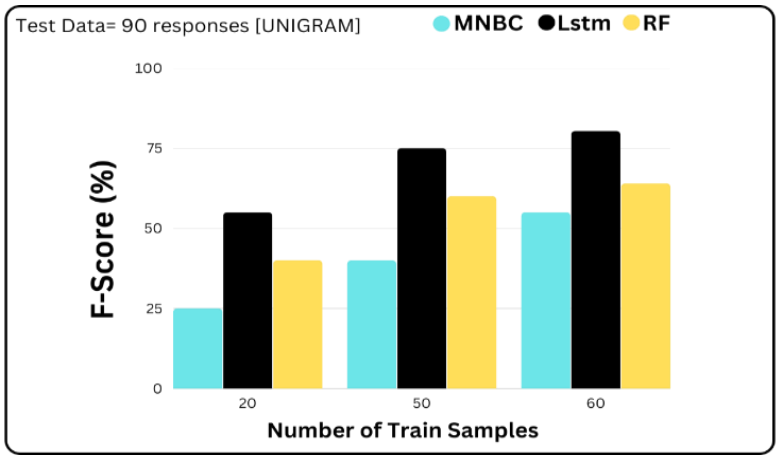

Figure 5: Shows a graph of the F Score for a Unigram versus the number of train samples.

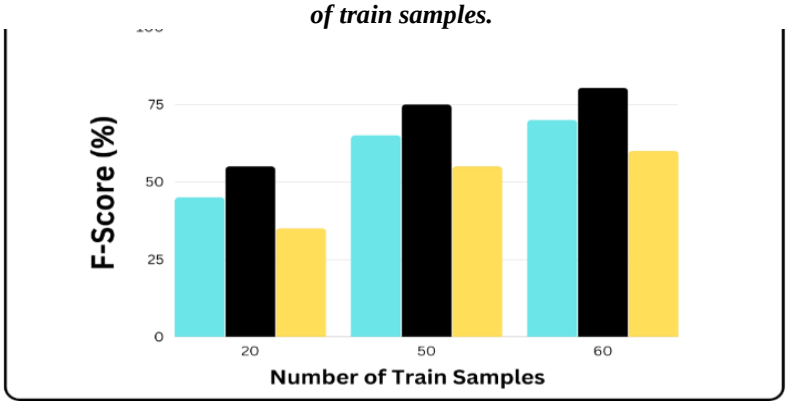

Figure 6: Shows a graph of Bigram's F Score as a function of the quantity of train samples.

Pakistan Journal of Engineering Technology and Science (PJETS)
Volume 11. Issue. 2, PP. 32-40, December 2023
https://doi.org/10.22555/pjets.v11i2.1039
E-ISSN 2224-2333
P-ISSN 2222-9930

Figure 7 illustrates the final working results of the sentiment analysis model you see that the student's feedbacks are provided to the Model and accurately predicts its sentiment label with an accuracy of 95.75%.

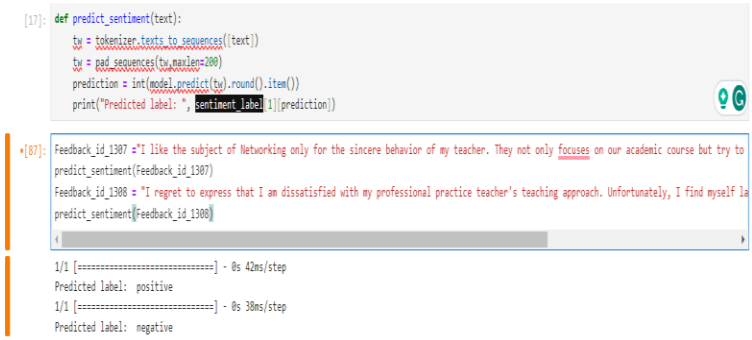

Figure 7: The final working results of Sentiment Analysis model.

IV. CONCLUSION AND FUTURE WORK

The negative aspect of the institutions' grading feedback is that it does not represent the students' sentiments fairly. Textual feedback fills this gap since it enables institutions to comprehend students' feelings and Sentiment analysis of this feedback is used it to improve the quality of education. The feedback is gathered and provided to the trained model in this study before the sentiments are categorized into positive, negative, and neutral. As per this study the results, of the long short-term memory algorithm provide higher accuracy than the Random Forest and Multinomial Naive Bayes Classifier algorithms in terms of accuracy. Moreover, the long short-term memory method performs improved than the other two methods with an accuracy of 95.75%. Future research in this area will likely focus on improving the model's accuracy by considering a large amount of training data.

REFERENCES

[1] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," IEEE Intell Syst, vol. 28, no. 2, pp. 15–21, 2013, doi: 10.1109/MIS.2013.30.

[2] S. Poria, E. Cambria, G. Winterstein, and G. Bin Huang, "Sentic patterns: Dependency-based rules for concept-level sentiment analysis," Knowl Based Syst, vol. 69, no. 1, pp. 45–63, 2014, doi: 10.1016/j.knosys.2014.05.005.

[3] M. Ali and A. Imdad, "Sentiment Summerization and Analysis of Sindhi Text," International Journal of Advanced Computer Science and Applications, vol. 8, no. 10, pp. 296–300, 2017, doi: 10.14569/ijacsa.2017.081038.

[4] D. D. Dsouza, Deepika, D. P. Nayak, E. J. Machado, and N. D. Adesh, "Sentimental analysis of student feedback using machine learning techniques," International Journal of Recent Technology and Engineering, vol. 8, no. 1 Special Issue 4, pp. 986–991, 2019.

[5] P. Ren, L. Yang, and F. Luo, "Automatic scoring of student feedback for teaching evaluation based on aspect-level sentiment analysis," Educ Inf Technol (Dordr), vol. 28, no. 1, pp. 797–814, Jan. 2023, doi: 10.1007/s10639-022-11151-z.

[6] Z. Nasim, Q. Rajput, and S. Haider, "Sentiment analysis of student feedback using machine learning and lexicon based approaches," International Conference on Research and Innovation in Information Systems, ICRIIS, pp. 1–6, 2017, doi: 10.1109/ICRIIS.2017.8002475.

[7] J. Yuan, Y. Wu, X. Lu, Y. Zhao, B. Qin, and T. Liu, "Recent advances in deep learning based sentiment analysis," Sci China Technol Sci, vol. 63, no. 10, pp. 1947–1970, Oct. 2020, doi: 10.1007/s11431-020-1634-3.

[8] P. Mata, J. X. Rita, A. Batista, and J. X. Rita, "Sentiment analysis – a literature review," Academy of Entrepreneurship Journal, vol. 27, no. SpecialIssue 2, pp. 1–10, 2021.

[9] B. K. Bhavitha, A. P. Rodrigues, and N. N. Chiplunkar, "Comparative study of machine learning techniques in sentimental analysis," Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2017, no. Icicct, pp. 216–221, 2017, doi: 10.1109/ICICCT.2017.7975191.

[10] K. S. Krishnaveni, R. R. Pai, and V. Iyer, "Faculty rating system based on student feedbacks using sentimental analysis," 2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2017, vol. 2017-Janua, pp. 1648–1653, 2017, doi: 10.1109/ICACCI.2017.8126079.

[11] R. M. Duwairi and I. Qarqaz, "Arabic sentiment analysis using supervised classification," in Proceedings - 2014 International Conference on Future Internet of Things and Cloud, FiCloud 2014, Institute of Electrical and Electronics Engineers Inc., Dec. 2014, pp. 579–583. doi: 10.1109/FiCloud.2014.100.

PJETS
E-ISSN 2224-2333
P-ISSN 2222-9930

Pakistan Journal of Engineering Technology and Science (PJETS)
Volume 11. Issue. 2, PP. 32-40, December 2023
https://doi.org/10.22555/pjets.v11i2.1039

[12] D. K. Dake and E. Gyimah, "Using sentiment analysis to evaluate qualitative students' responses," Educ Inf Technol (Dordr), vol. 28, no. 4, pp. 4629–4647, Apr. 2023, doi: 10.1007/s10639-022-11349-1.
[13] J. Zhou and J. min Ye, "Sentiment analysis in education research: a review of journal publications," Interactive Learning Environments. Routledge, 2020. doi: 10.1080/10494820.2020.1826985.
[14] M. Li, L. Chen, J. Zhao, and Q. Li, "Sentiment analysis of Chinese stock reviews based on BERT model", doi: 10.1007/s10489-020-02101-8/Published.
[15] M. Baygin, "Classification of Text Documents based on Naive Bayes using N-Gram Features," 2018 International Conference on Artificial Intelligence and Data Processing, IDAP 2018, pp. 1–5, 2019, doi: 10.1109/IDAP.2018.8620853.
[16] X. Fang and J. Zhan, "Sentiment analysis using product review data," J Big Data, vol. 2, no. 1, 2015, doi: 10.1186/s40537-015-0015-2.
[17] F. Rustam, M. Khalid, W. Aslam, V. Rupapara, A. Mehmood, and G. S. Choi, "A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis," PLoS One, vol. 16, no. 2, Feb. 2021, doi: 10.1371/journal.pone.0245909.
[18] M. Mujahid et al., "Sentiment analysis and topic modeling on tweets about online education during covid-19," Applied Sciences (Switzerland), vol. 11, no. 18, Sep. 2021, doi: 10.3390/app11188438.
[19] W. Liao, B. Zeng, X. Yin, and P. Wei, "An improved aspect-category sentiment analysis model for text sentiment analysis based on RoBERTa," Applied Intelligence, vol. 51, no. 6, pp. 3522–3533, Jun. 2021, doi: 10.1007/s10489-020-01964-1.
[20] M. Zhou, N. Duan, S. Liu, and H. Y. Shum, "Progress in Neural NLP: Modeling, Learning, and Reasoning," Engineering, vol. 6, no. 3, pp. 275–290, 2020, doi: 10.1016/j.eng.2019.12.014.
[21] G. Qiang, "Research and improvement for feature selection on naive bayes text classifier," Proceedings of the 2010 2nd International Conference on Future Computer and Communication, ICFCC 2010, vol. 2, pp. 156–159, 2010, doi: 10.1109/ICFCC.2010.5497362.
[22] S. Xu, Y. Li, and Z. Wang, "Bayesian multinomial naïve bayes classifier to text classification," Lecture Notes in Electrical Engineering, vol. 448, no. 15, pp. 347–352, 2017, doi: 10.1007/978-981-10-5041-1_57.
[23] N. Sharma and M. Singh, "Modifying Naive Bayes classifier for multinomial text classification," 2016 International Conference on Recent Advances and Innovations in Engineering, ICRAIE 2016, 2016, doi: 10.1109/ICRAIE.2016.7939519.
[24] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 67, no. 1, pp. 93–104, 2012, doi: 10.1016/j.isprsjprs.2011.11.002.
[25] M. Junker, R. Hoch, and A. Dengel, "On the evaluation of document analysis components by recall, precision, and accuracy," Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, pp. 717–720, 1999, doi: 10.1109/ICDAR.1999.791887.
[26] B. Marie, A. Fujita, and R. Rubino, "Scientific credibility of machine translation research: A meta-evaluation of 769 papers," ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference, pp. 7297–7306, 2021, doi: 10.18653/v1/2021.acl-long.566.