Published in AJSE, Vol:22, Issue: 2
Received on 9th July 2023
Revised on 20th August 2023
Accepted on 22th August 2023

AIUB JOURNAL OF SCIENCE AND ENGINEERING
ISSN: 1608 – 3679 (print) 2520 – 4890 (Online)

A comprehensive dataset for aspect-based sentiment analysis in evaluating teacher performance

Abhijit Bhowmik, Noorhuzaimi Mohd Noor, M. Saef Ullah Miah, Md. Mazid-Ul-Haque, and Debajyoti Karmaker

***Abstract*—Teacher performance evaluation is an essential task in the field of education. In recent years, aspect-based sentiment analysis (ABSA) has emerged as a promising technique for evaluating teaching performance by providing a more nuanced analysis of student evaluations. This article presents a novel approach for creating a large-scale dataset for ABSA of teacher performance evaluation. The dataset was constructed by collecting student feedback from American International University-Bangladesh and then labeled by undergraduate-level students into three sentiment classes: positive, negative, and neutral. The dataset was carefully cleaned and preprocessed to ensure data quality and consistency. The final dataset contains over 2,000,000 student feedback instances related to teacher performance, making it one of the largest datasets for ABSA of teacher performance evaluation. This dataset can be used to develop and evaluate ABSA models for teacher performance evaluation, ultimately leading to better feedback and improvement for educators. The results of this study demonstrate the usefulness and effectiveness of ABSA in evaluating teacher performance and highlight the importance of creating high-quality datasets for this task.**

***Index Terms*—Sentiment analysis dataset, Aspect based sentiment analysis, NLP, Data processing, Data preparation**

I. INTRODUCTION

As educational institutions strive to enhance the quality of teaching, and the evaluation of teacher performance using objective and reliable measures has become increasingly important [1]. One such measure is sentiment analysis, which employs natural language processing and machine learning algorithms to examine the opinions and emotions expressed in text[2], [3].

Aspect-based sentiment analysis (ABSA) is a subfield of sentiment analysis that focuses on identifying and analyzing the sentiment associated with specific aspects or features of a product or service [4]. In the context of teacher performance evaluation, ABSA can be utilized to identify the strengths and weaknesses of teachers in particular areas such as classroom management, student engagement, and content delivery.

However, the lack of suitable datasets presents a significant challenge in implementing ABSA for teacher performance evaluation [5]. A high-quality dataset is crucial for training machine learning models that can accurately identify and classify sentiments associated with specific aspects of teaching. The accuracy of ABSA models is likely low without a high-quality dataset, leading to unreliable evaluations of teacher performance [6], [7], [8], [9].

A good dataset for ABSA in teacher performance evaluation has many advantages. Firstly, it allows for more objective and reliable evaluations of teacher performance, reducing the influence of subjective biases that may be present in traditional evaluation methods. Secondly, it permits more detailed and nuanced evaluations of specific aspects of teaching, providing valuable insights for enhancing teaching practices. Finally, it facilitates the development of more sophisticated ABSA models that can adapt to the unique characteristics of various educational contexts.

Conversely, the absence of a suitable dataset for ABSA in teacher performance evaluation can lead to inaccurate and unreliable evaluations of teacher performance, potentially resulting in unjust or inequitable outcomes. Additionally, the lack of a high-quality dataset can impede the development of ABSA models for teacher performance evaluation, limiting the potential of sentiment analysis as a tool for improving the quality of teaching.

This research article addresses the need for a high-quality dataset for ABSA in teacher performance evaluation by presenting a novel dataset specifically designed for this purpose. We believe this dataset will benefit educational institutions, policymakers, and educational researchers, as it provides a more reliable and objective approach to evaluating teacher performance. By providing a more accurate assessment of teaching practices, we believe our research will improve the quality of education and ultimately benefit society.

This dataset is made publicly available for future research purposes and can be found at URL: https://doi.org/10.17632/b2yhc95rnx.

II. LITERATURE REVIEW

The use of sentiment analysis for evaluating teacher performance is a relatively new research area, and as such, there is limited literature on the subject. However, sentiment analysis has been widely used in other domains such as product reviews

Abhijit Bhowmik is with the Faculty of Computing, College of Computing and Applied Sciences, Universiti Malaysia Pahang Al-Sultan Abdullah (UMPSA), 26600, Pekan, Malaysia, And Faculty of Science and Technology, Department of Computer Science, American International University-Bangladesh (AIUB), 1229, Dhaka, Bangladesh (e-mail: ovi775@gmail.com).
Noorhuzaimi Mohd Noor is with the Faculty of Computing, College of Computing and Applied Sciences, Universiti Malaysia Pahang Al-Sultan Abdullah (UMPSA), 26600, Pekan, Malaysia (e-mail: nhuzaimi@ump.edu.my).
M. Saef Ullah Miah is with the Faculty of Science and Technology, Department of Computer Science, American International University-Bangladesh (AIUB), 1229, Dhaka, Bangladesh (e-mail: md.saefullah@gmail.com).
Md. Mazid-Ul-Haque is with the Faculty of Science and Technology, Department of Computer Science, American International University-Bangladesh (AIUB), 1229, Dhaka, Bangladesh (e-mail: mazid@aiub.edu).
Debajyoti Karmaker is with the Faculty of Science and Technology, Department of Computer Science, American International University-Bangladesh (AIUB), Research Associate, The University of New South Wales (UNSW). (e-mail: d.karmaker@aiub.edu).

[10], social media analysis [11], and customer feedback analysis [12]. Therefore, existing literature on sentiment analysis can provide insights into the potential benefits and limitations of using sentiment analysis for teacher performance evaluation.

One of the key advantages of using sentiment analysis for teacher performance evaluation is its ability to provide more objective and reliable evaluations of teaching practices. Traditional evaluation methods such as student surveys or peer evaluations are often biased and can be influenced by student expectations, personal relationships, or personal preferences [13]. In contrast, sentiment analysis provides a more data-driven approach to evaluating teacher performance, which reduces the impact of subjective biases [14].

A review of existing literature also reveals the importance of having a good dataset for sentiment analysis [15], [16], [17], [18], [19], [6]. A good dataset is critical for training machine learning models that can accurately identify and classify sentiments associated with specific aspects of teaching. Existing research in sentiment analysis has shown that the accuracy of machine learning models is highly dependent on the quality and quantity of the dataset used for training.

However, despite the potential benefits of using sentiment analysis for teacher performance evaluation, several limitations and challenges must be considered. Firstly, sentiment analysis is primarily based on analyzing text data, which may not capture all aspects of teaching performance, such as non-verbal communication, body language, or classroom dynamics. Secondly, contextual factors such as students' cultural background, the subject matter being taught, and the teaching methods used can be influenced sentiment analysis. Therefore, it is important to carefully select the aspects of teaching that are analyzed using sentiment analysis and to ensure that the results are interpreted in the appropriate context.

Additionally, the lack of a suitable dataset for sentiment analysis is a major challenge that needs to be addressed. Most existing datasets in sentiment analysis are focused on product reviews or social media analysis, and there are few datasets specifically designed for analyzing teacher performance. Therefore, there is a desperate need to develop new datasets tailored to the unique characteristics of teacher performance evaluation.

The literature review suggests that sentiment analysis has the potential to provide more objective and reliable evaluations of teacher performance. Still, the dataset's quality for training machine learning models is critical to its success. Furthermore, the limitations and challenges associated with sentiment analysis should be carefully considered when designing evaluation methods. The development of a novel dataset for aspect-based sentiment analysis for teacher performance evaluation, as presented in this article, can provide a valuable contribution to the existing literature and help overcome some of the limitations and challenges associated with sentiment analysis in teacher performance evaluation.

III. Dataset Description and Development

This research article presents a novel dataset for aspect-based sentiment analysis for teacher performance evaluation. The dataset was collected using manual and automated methods to ensure high quality and accuracy.

The dataset consists of reviews of teaching performance written by students from various courses from different disciplines, namely, engineering, business, sociology and computer science. The reviews were collected over the Summer 2003-2004 semester to the Spring 2021-2022 semester, which is around 18 years and covers various teaching aspects such as classroom management, student engagement, and content delivery.

A three-step process was used to filter out irrelevant or low-quality reviews to ensure the dataset's quality. Firstly, reviews were automatically filtered based on the presence of certain keywords related to teaching performance. Secondly, human annotators manually reviewed the filtered reviews to eliminate any remaining irrelevant or low-quality reviews. Finally, a second team of human annotators validated the sentiment assigned by the deep learning-based sentiment analysis model.

Benchmarking datasets is essential for orienting machine learning communities' goals and measuring progress in the field [20], [21], [22]. However, the near-exclusive focus on boosting benchmark metrics has been criticized from various angles. Likewise, the current benchmarking culture has been blamed for stifling the development of innovative ideas [23], [24]. Datasets that support machine learning are frequently utilized, shared, and reused with limited visibility into the decision processes that lead to their formation. As artificial intelligence systems become more prevalent in high-stakes jobs, system development, and deployment processes must evolve to meet the real repercussions of how model development data is created and used in practice [25]. This includes improved transparency regarding data and accountability for data-development decisions.

The above-mentioned concerns are considered when developing the dataset for this research. Figure 1 shows the dataset development process with the necessary steps discussed elaborately in the following subsections.

A. Data Collection

Collecting data is a crucial challenge for machine learning and a widely discussed topic in many communities. This concern has recently become more critical for two main reasons [26]. First, with the increasing use of machine learning, new emerging applications may not have enough labeled data available. Second, unlike conventional machine learning, deep learning algorithms can automatically create features, which saves on feature engineering but may require more labeled data. It is worth noting that due to the importance of managing large volumes of data, research on data collection also arises from the data management community in addition to the machine learning, natural language processing, and computer vision fields. The primary aspects of data collection include gathering, categorizing, and improving new data or models. Data quality is a significant concern when collecting data, as unstructured data is often acquired without the necessary details for problem diagnosis [27]. This study collects data from the virtual university expert system (VUES) of American International University-Bangladesh (AIUB), including students'

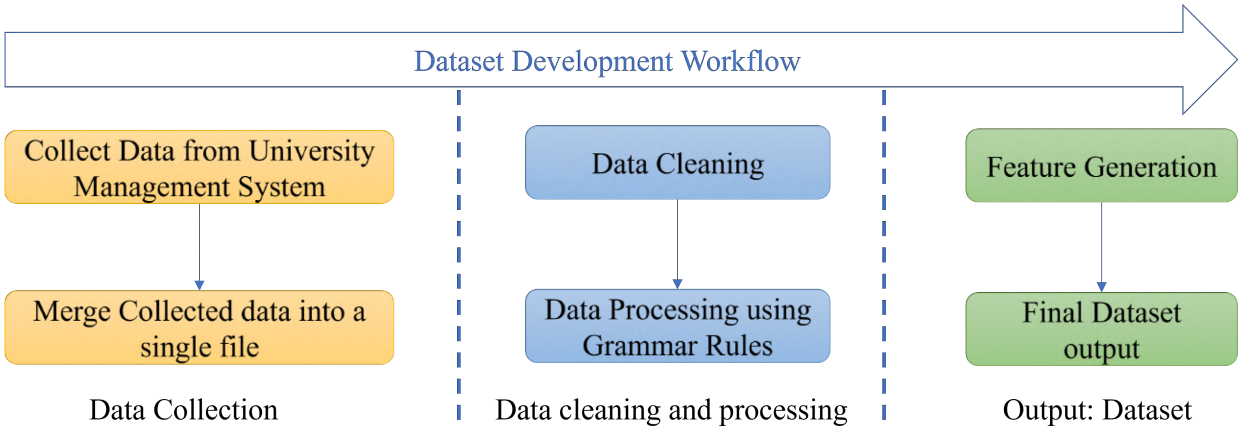

Fig. 1: Dataset development methodology

ratings and comments. Each semester's data is contained in a single Excel file, and after collecting all the semesters' data, they are combined into a single tab-separated value (TSV) formatted file. The original data collected from the system contains 22,04,523 rows, which are then cleaned and processed to ensure data quality in the following phase. Figure 15 in Appendix shows the structure of the data contained in each single Microsoft Excel file. The figure depicts an example of raw data consisting of several rows arranged in a table format. Each row contains information related to student feedback and the course they have taken. The columns in the table include Student Comments, Rating, Course ID, Offered Course ID, Course Name, Section, and Semester. The Student Comments column contains comments provided by the students, while the Rating column contains their ratings. The Course ID and Offered Course ID columns contain unique identifiers for the course and its offerings. The Course Name column provides the name of the course, while the Section column contains the section number associated with the course. Finally, the Semester column indicates the semester in which the course was offered. The information in columns Course ID and Offered Course ID are removed for anonymity.

B. Data Processing and Cleaning

Data cleaning is crucial in ensuring the dataset is free of wrong or erroneous data because it is the first stage of any machine learning activity and one of the most critical procedures in data analysis. The model's performance is determined by the data used to train it, making data preparation a crucial step in developing the classification model. Reducing the noise in the data and removing the useless data leads to the best possible outcome. First, the data set is examined for null or blank values, and the required actions are taken. The columns used to train the model are cleaned by removing any punctuation, HTML tags, special characters, numbers, and extra whitespace. As we are utilising transformer-based pre-trained models to train the classifier, we are avoiding typical text pre-processing approaches such as stop word removal, stemming, and lemmatization to preserve the semantic contents of the reviews. Before feeding the input to pre-trained models, the raw data is translated to an appropriate format by tokenizing each sentence using tokenizers unique to the pre-trained model. Algorithm 1 presents a step-by-step process to clean the dataset.

Algorithm 1 Data Cleaning Algorithm

```
Require: dataset in tsv format
Ensure: cleaned dataset
Mount the dataset on Google Drive
Import pandas library
Read the tsv file using pd.read_csv(dataset_dir, sep='
t')
Check the first 2 rows and the size of the dataframe using
  head(2) and shape functions
Delete all the columns except StudentsComments and
  Rating
Use strip() function to delete unwanted spaces from the
  beginning and end of the StudentComments column
Use replace() function to remove multiple special char-
  acters like space between words, dot (.), comma (,) to a
  single character, and to remove all the special characters
  like !, -, ?, @, *, #, $, %
Use replace() function to make the empty call null and
  delete that row using dropna()
Convert all the StudentComments to lower case using
  lower() function
Delete rows containing "no comment" and/or "no com-
  ments"
Count the word frequency of all the StudentComments
  and store it in a new column named totalwords
for each row in the dataframe do
    if it is a unigram then
        if the word is not in the wordnet then
            Delete the row
        end if
    end if
end for
return cleaned dataset
```

The algorithm begins by mounting the dataset on Google

Drive and importing the pandas library, which is a popular library for data manipulation and analysis in Python. It then reads the TSV file using the pandas function *read_csv()* and checks the first 2 rows and the size of the dataframe using *head(2)* and *shape()* functions.

The next step is to delete all the columns except for "StudentsComments" and "Rating". It then uses the *strip()* function to delete unwanted spaces from the beginning and end of the"StudentsComments" column. After that, it uses the *replace()* function to remove multiple special characters like space between words, dot (.), comma (,), etc., and to remove all the special characters like !, -, ?, @, *, #, $, %.

The algorithm then deletes any rows that have empty comments by replacing the empty cell with null values and deleting that row using the *dropna()* function. It also converts all the comments to lowercase using the *lower()* function.

It then deletes rows that contain the phrases "no comment" and/or "no comments". The algorithm then counts the frequency of words in each comment and stores it in a new column called "totalwords".

Next, the algorithm loops through each row in the dataframe and checks if the comment contains a single word (unigram). If it is a unigram, it checks if the word is in the WordNet dictionary. If it is not in the dictionary, the algorithm deletes the row.

Finally, the cleaned dataset is returned. Overall, this algorithm performs a series of data cleaning operations on a dataset to improve its quality for analysis.

The above-mentioned steps are followed to perform data preprocessing and cleaning. A significant amount of noise reduction can be observed here. Before cleaning the number of rows was 2,204,522, and after cleaning the number of rows are 2,007,747 with 381,005 distinct comments which are about 19% of the total data. The count distinct rating is 401. With this cleaned data, the research advances to the next phases. Table I provides a summary of the data processing steps taken and the resulting data sizes after each step. The table includes two columns, one for the data processing step and one for the resulting data size.

TABLE I: Summary of Data Processing Steps and Data Sizes

Data Processing Step	Data Size
Raw Data	2204522
After Cleaning Null, Blank Values, and Special Characters	2184387
After Cleaning Comments Like No Comment/(s)	2120997
After Removing Meaningless Unigrams	2007747
Final Data Size	2007747

C. Dataset Preparation

The next step in a machine learning project is data preparation, also known as data curation. This can take a long time, especially for huge data sets, and involves dealing with duplicate data, missing data, and other formatting difficulties. Model training and data storage can be done locally on hardware or remotely using cloud computing services.

For this study, a new data set is curated from the cleaned data by checking all rows for sentence and word length. If the word length is one or two, it is checked whether they are present in the dictionary or not. If they are present, they are kept or the line is deleted. Then all ratings are rounded as floating point scores and are considered as items such as 1, 2, 3, 4, and 5, where 5 is counted as the best quality and 1 as the lowest quality. After that, sentiments are devised from the rating scores. If the score is less than 3 it is considered negative, if it is 3 then considered neutral and a score greater than 3 is considered a positive sentiment. After that, a pre-trained deep learning-based model namely, cardiffnlp/twitter-roberta-base-sentiment [28] model is deployed to label all the StudentCommnets' sentiment and the sentiment value for each comment is stored alongside the rating devised sentiment. After that, the subjectivity and objectivity are calculated for each comment using the TextBlob [29] library and stored in a separate column. Next, the token count is performed on the student comments and the number of tokens or words is stored in another separate column. After that, the matching process between the two sentiment columns is done by a Python script which checks if both columns' values are the same or not. If the values are the same it is labeled as true else as fake. After that the group of human annotators validated the true and fake labels if they are true or not. Finally, the IAA scores are calculated using Fleiss' Kappa for the labeled dataset. Algorithm 2 presents the steps followed to prepare the final dataset. Figure 16 in Appendix shows the final structure of the dataset. The figure displays an example of the final curated data in a tabular format. It has several columns, including StudentComments, Rating, totalwords, Sentiment, sent_pretrained, subjectivity, subj-score, and isSame. The StudentComments column contains comments written by students about a particular course. The Rating column shows the corresponding rating given by the student for the same course. The totalwords column indicates the total number of words in each student comment. The Sentiment column shows the sentiment of the comment, which is either negative, neutral, or positive. The sent_pretrained column indicates the sentiment of the comment predicted by a pre-trained deep learning-based model. The subjectivity column shows the degree of subjectivity in each comment, and the subj-score column displays the corresponding score for the degree of subjectivity. Finally, the isSame column shows whether the sentiment values in the Sentiment and sent_pretrained columns match or not, which is validated by human annotators.

The following subsection discusses the exploratory data analysis on the prepared dataset for this study which helps to understand the data and identify patterns, relationships, and anomalies.

D. Dataset Overview

Before cleaning, the dataset consisted of 2204522 entries with 2 properties named studentscomments and rating. After cleaning the null, blank values, and unnecessary special characters like multiple spaces ( ), dots (.), commas(,), exclamatory signs (!), hyphens (-), question marks (?), at the rate (@), asteroids (*), hash (#), dollar signs ($), percentage (%) from students comments and rating, the dataset had 2184387 entries.

Algorithm 2 Data Preparation Algorithm

Require: dataset in tsv format
Ensure: processed dataset
Mount the dataset on Google Drive
Import pandas, TextBlob, and transformers libraries
Read the tsv file using pd.read_csv (dataset_dir, sep='tab')
Delete all the columns except StudentsComments and Rating
for each row in the dataframe **do**
    if the length of the word is one or two **then**
        if the word is not in the dictionary **then**
            Delete the row
        end if
    end if
end for
Round all ratings as floating point scores and consider them as items such as 1, 2, 3, 4, 5, where 5 is considered as the best quality and 1 as the lowest quality
Devises sentiment from the rating scores, where scores less than 3 are considered negative, 3 as neutral, and greater than 3 as positive sentiment
Deploy pre-trained deep learning-based model named cardiffnlp/twitter-roberta-base-sentiment [28] to label all the StudentComments' sentiment and store the sentiment value for each comment alongside the rating devised sentiment
Calculate the subjectivity and objectivity for each comment using the TextBlob library and store them in a separate column
Perform token count on the student comments and store the number of tokens or words in another separate column
Match the two sentiment columns using a Python script that checks if both columns have the same value or not
Label the match as true if the values are the same, otherwise as fake
Validate the true and fake labels by a group of human annotators
Calculate the IAA scores using Fleiss' Kappa for the labelled dataset
return processed dataset

After cleaning comments like no comment/(s), the dataset became entries with 2120997. After checking if the comments had meaningful unigram or not. After removing the meaningless unigrams, the dataset had 2007747 entries. Fig. 15 and Fig. 16 in Appendix show the structure of the raw data and curated final TPE dataset respectively.

E. Exploratory Data Analysis

Exploratory data analysis is significant since it contributes to the optimization of data. By integrating data analysis into the model, researchers can find more efficient ways to conduct their operations and store vast volumes of data. The goal of data analysis is to provide accurate and trustworthy data. And filtering missing, uncompleted, and empty data can give more accuracy in modeling. For these purposes, in this section, the dataset has been explored deeply to get a better understanding of the data. The distribution of different feature counts of ratings by rating number in the cleaned dataset is shown in Figure 2.

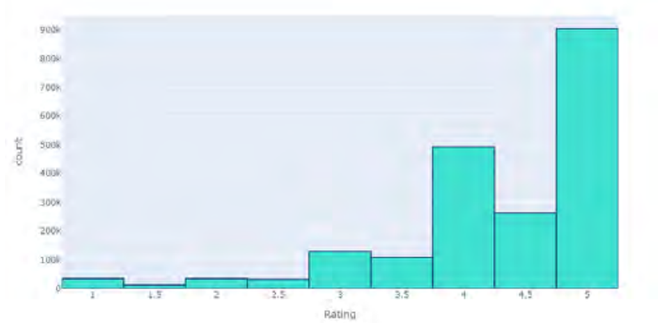

Fig. 2: Total counts of rating in the dataset

An overview of the words used frequently in the student comments is shown in a word cloud representation in Figure 3. As per the figure, the most commonly used words in the students' comments feature are good teacher, good, good faculty, best teacher, best, teacher, friendly, excellent teacher, and nice teacher. The ratings and comments are aligned, and students have been given proper comments and ratings.

Fig. 3: Commonly used words in the student's comments feature

The distribution of the number of words per rating in student comments text is shown in Figure 4. In this figure, it can be observed that the number of words per rating is slightly increasing with the increase of the rating point.

The distribution of sentiment classes found in the dataset using the rating feature is shown in Figure 5. From this figure, it can be seen that the sentiment classes are categorized into three classes, positive, neutral and negative. The positive class has the highest weight, followed by neutral and negative.

Word clouds for different classes (cleaning and pre-trained sentiment) where frequent phrases can be observed for negative sentiments categorized by ratings are shown in Figure 6. As per this figure, frequently used words for negative reviews

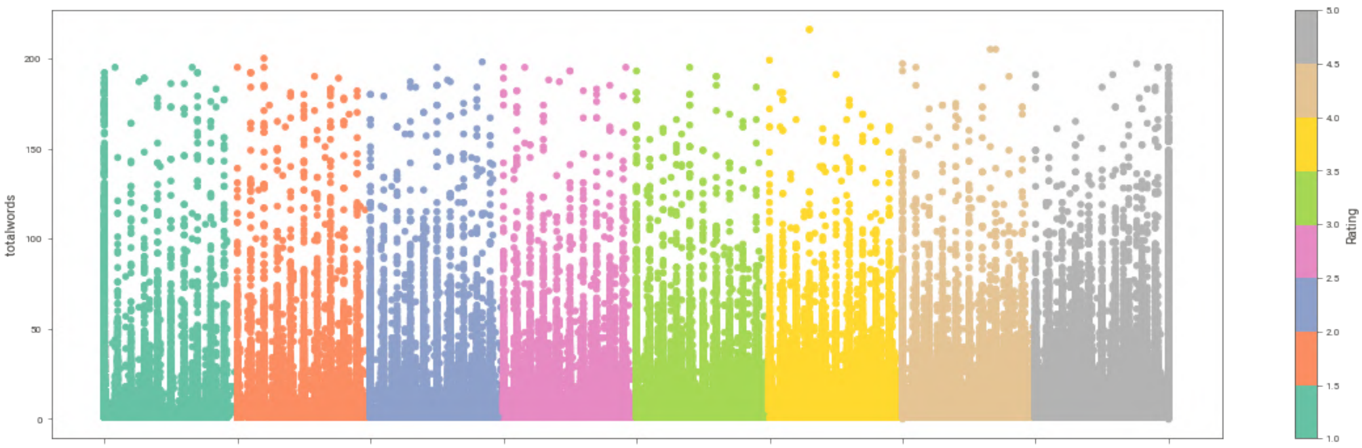

Fig. 4: Number of words per rating

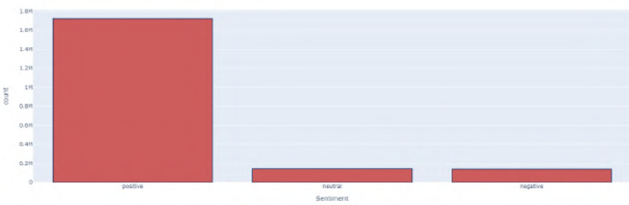

Fig. 5: Distribution of sentiments categorized by rating

were good good, good, good teacher, and understand, which means students were not giving proper comments with their ratings. They gave negative or low ratings but wrote positive comments.

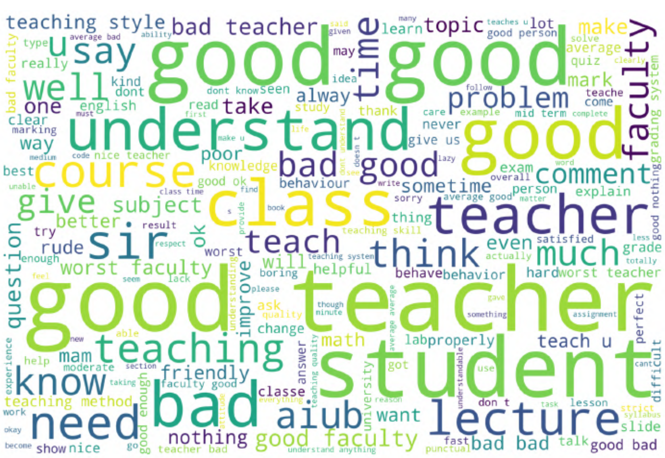

Fig. 6: Frequent phrases observed for negative sentiments by rating

Frequent phrases observed for negative sentiments categorized by ratings are shown in Figure 7. Same as the students who provided negative comments, students who commented neutrally have also given positive comments, but the rating provided can be categorized as neutral or mid-level.

Finally, the dataset had 2007747 entries with 6 new columns named totalwords, sentiment, sent_pretrained, subjectivity,

Fig. 7: Frequent phrases observed for neutral sentiments by rating

subj-score, and is same means a total of 8 columns.

In the trained dataset, sent_pretained column had students comments type such as positive, negative, and neutral, and after creating word clouds for positive, negative, and neutral comments, we saw,

Positive comments had words like good good, best faculty, best teacher, good teacher, understand. These are almost the same as the previous positive word cloud. Still, in new negative and neutral word clouds, some new words have been seen, such as bad, answer, need, nothing, question, mark, teaching, style, slide, problem in negative, and moderate, understanding, difficult, overall good, example, given, good enough in neutral. These words are different from previous word clouds and are aligned with the sent_pretained means comments type and rating.

The significant findings through the analysis of data are presented throughout this subsection. These outcomes are crucial for the model development process through which fraud reviews and ratings will be detected. The following section presents the statistical analysis of the dataset.

F. Statistical Analysis

1) Descriptive Statistics: Table II represents descriptive statistics for the numerical columns ”Rating,” ”totalwords,” and ”subj-score” of the dataset.

TABLE II: Descriptive Statistics for Numerical Columns

Measure	Rating	totalwords	subj-score
Mean	4.28795	4.76135	0.559776
Median	4.55	2	0.6
Mode	5	1	0.6
Standard Deviation	0.867526	8.17894	0.248453
Range	4	216	1
25th Percentile (Q1)	4	1	0.5
50th Percentile (Median)	4.55	2	0.6
75th Percentile (Q3)	5	5	0.6

The table presents a comprehensive overview of the descriptive statistics calculated for the dataset's numerical columns. Descriptive statistics provide valuable insights into the data's central tendency, spread, and distribution.

The mean, also known as the average, measures the central value for each column. For instance, the mean rating is approximately 4.29, indicating that the average rating given by students is around 4.29. Similarly, the mean total word count is approximately 4.76, suggesting that the average comment length is about 4.76. The mean subjectivity score is approximately 0.56, which provides insight into the average subjectivity level of the comments.

The median is another measure of central tendency representing the middle value when the data is sorted in ascending order. For example, the median rating is 4.55, indicating that half of the ratings fall below 4.55 and half are above it. The median total word count is 2, meaning half of the comments have a word count less than or equal to 2, and the other half have word counts greater than or equal to 2. Similarly, the median subjectivity score is 0.6, reflecting the middle value of the subjectivity scores.

The mode, the most frequently occurring value, provides insights into the most common values within each column. For instance, the mode for ”Rating” and ”subj-score” is 5, suggesting that 5 is the most common rating and subjectivity score among the entries. For ”totalwords,” the mode is 1, indicating that a word count of 1 is the most prevalent.

The standard deviation measures the dispersion of data points around the mean, providing a sense of how much the values deviate from the average. A higher standard deviation signifies greater variability in the data. For instance, the standard deviation for ”Rating” is about 0.87, indicating that the ratings are spread around the mean of 4.29 with a certain degree of variability. Similarly, the standard deviation for ”totalwords” is approximately 8.18, indicating a wider word count spread around the mean of 4.76. The standard deviation for ”subj-score” is approximately 0.25, suggesting less variability in subjectivity scores around the mean of 0.56.

The range represents the difference between each column's maximum and minimum values. For example, the range for ”Rating” is 4 (from 1 to 5), indicating the full spread of ratings in the dataset. For ”totalwords,” the range is 216 (the maximum word count is 216 and the minimum is 1), reflecting the broad variation in comment lengths. For the ”subj-score,” the range is 1 (ranging from 0 to 1), showing the complete coverage of subjectivity scores.

The quartiles (25^{th}, 50^{th}, and 75^{th} percentiles) divide the data into four equal parts, providing insights into the distribution across the dataset. The 25^{th} percentile (Q1) for ”Rating” is 4, suggesting that 25% of the ratings are 4 or below. Similarly, the 25^{th} percentile for “totalwords” is 1, indicating that a quarter of the comments have a word count of 1 or less. The 25^{th} percentile for “subj-score” is 0.5, showing the lower 25% of subjectivity scores. The 50^{th} percentile (median) values are the same as discussed earlier. The 75^{th} percentile (Q3) values represent the cutoff below which 75% of the data falls. For example, the 75^{th} percentile for ”Rating” is 5, indicating that 75% of the ratings are 5 or below, and for “totalwords,” it is 5, revealing that 75% of the comments have a word count of 5 or less. Finally, the 75^{th} percentile for “subj-score” is 0.6, showing the upper 75% of subjectivity scores.

2) Correlation Analysis: For this dataset the Pearson Correlation analysis is performed on the numerical columns of the dataset. Table III shows the correlation matrix for all the numerical columns.

TABLE III: Correlation Matrix

	Rating	totalwords	subj-score
Rating	1.000000	-0.048633	0.113343
totalwords	-0.048633	1.000000	-0.022737
subj-score	0.113343	-0.022737	1.000000

This table shows the pairwise Pearson correlation coefficients between three numerical columns: “Rating,” “totalwords,” and “subj-score.” The values in each cell represent the strength and direction of the linear relationship between the corresponding pair of columns. The table indicates a weak negative correlation of approximately -0.0486 between “Rating” and “totalwords”. This suggests that as the ratings increase, there is a slight tendency for the total word count in the comments to decrease and vice versa. Furthermore, the table displays a weak positive correlation of approximately 0.1133 between “Rating” and “subj-score.” This indicates that higher ratings are slightly associated with higher subjectivity scores in the comments. Again, the correlation is relatively weak, and individual cases may not always follow this pattern. Lastly, the table reveals a negligible negative correlation of approximately -0.0227 between “totalwords” and “subj-score”. This implies a slight tendency for longer comments to have lower subjectivity scores. The visual representation of the Pearson correlation heatmap is shown in Fig. 8.

The correlation matrix indicates no substantial linear relationships among the three numerical columns. The coefficients are all relatively small, signifying weak or negligible correlations. This suggests that changes in one column do not consistently result in predictable changes in the other, indicating a relatively independent nature of these numerical variables.

3) Frequency Analysis: The frequency analysis is done for the categorical values. Table IV shows the frequency for

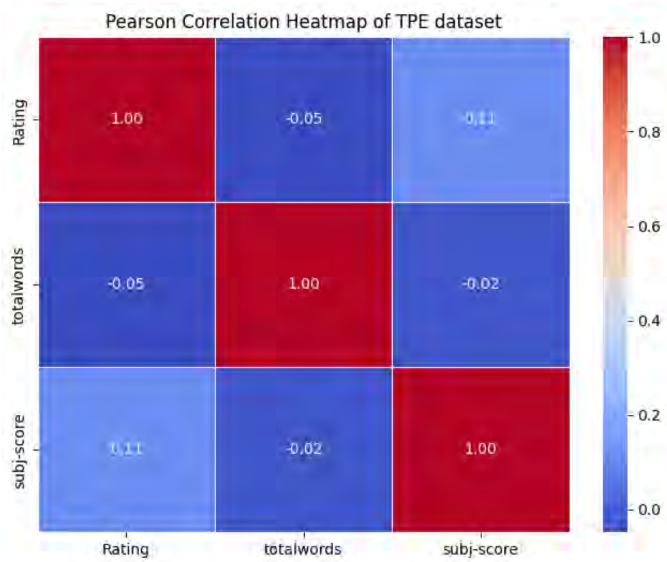

Fig. 8: Pearson correlation heatmap of TPE dataset for the numerical columns.

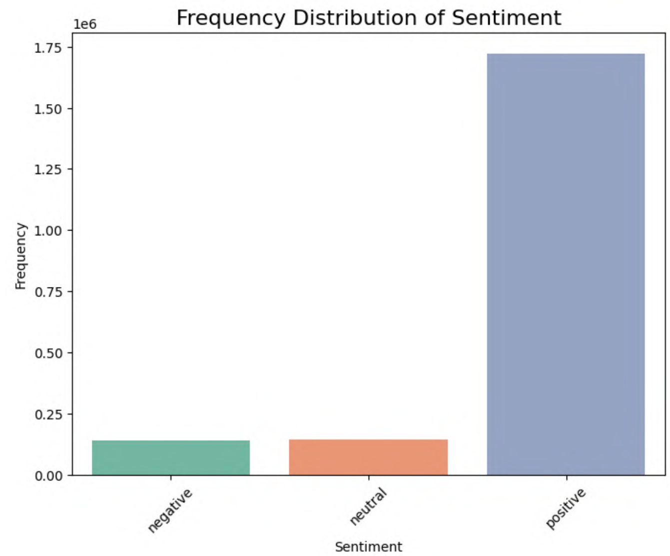

Fig. 9: Frequency plot for sentiment column

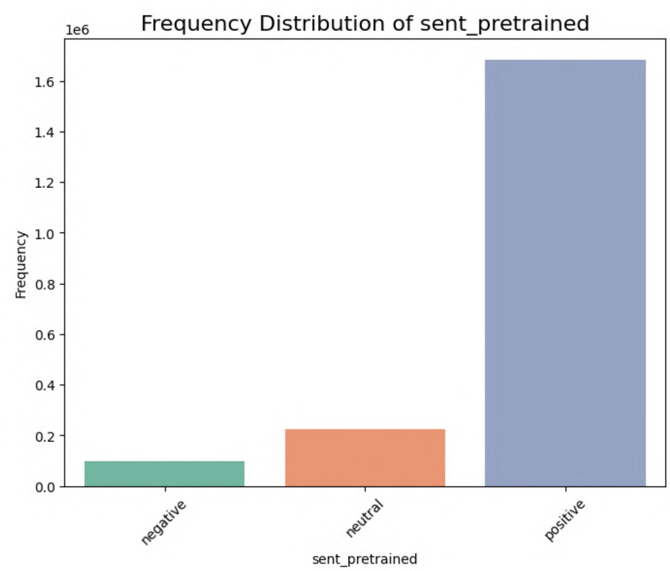

Fig. 10: Frequency plot for sentiment_pretrained column

the Sentiment column. Table V shows the frequency for the sent_pretrained column and Table VI shows the frequency for the subjectivity column.

TABLE IV: Frequency Table for 'Sentiment'

Sentiment	Count
positive	1,722,039
neutral	144,505
negative	141,203

TABLE V: Frequency Table for 'sent_pretrained'

sent_pretrained	Count
positive	1,684,600
neutral	223,442
negative	99,705

TABLE VI: Frequency Table for 'subjectivity'

subjectivity	Count
subjective	1,551,132
objective	456,615

4) ANOVA Test: The ANOVA test compares the means of two or more groups to determine if there are any significant differences between them. In the TPE dataset, there is a categorical column "Sentiment" representing different groups and numerical columns "Rating" and "totalwords" that can be compared among the groups. Table VII shows the ANOVA test results.

The table presents the results of the ANOVA test for three numerical variables: "Rating," "totalwords", and "subj-score." Each row represents a variable, and the columns display the F-Statistic and P-Value obtained from the ANOVA test. The F-Statistic measures the variation between group means relative to the variation within groups. It is used to assess whether significant differences exist in the means of the numerical variable among the groups. The P-Value indicates the probability of obtaining the observed F-Statistic, assuming the null hypothesis is true (i.e., no significant differences between the group means). A small P-Value (usually less than 0.05) suggests that the observed differences are unlikely to occur by chance, leading to the rejection of the null hypothesis.

In this case, for all three variables (Rating, totalwords, and subj-score), the P-Values are 0.00, indicating significant differences in the means of these numerical variables among the groups. Therefore, the null hypothesis is rejected, and it

TABLE VII: ANOVA Test Results

Variable	F-Statistic	P-Value
Rating	2351757.35	0.00
totalwords	9909.33	0.00
subj-score	10445.69	0.00

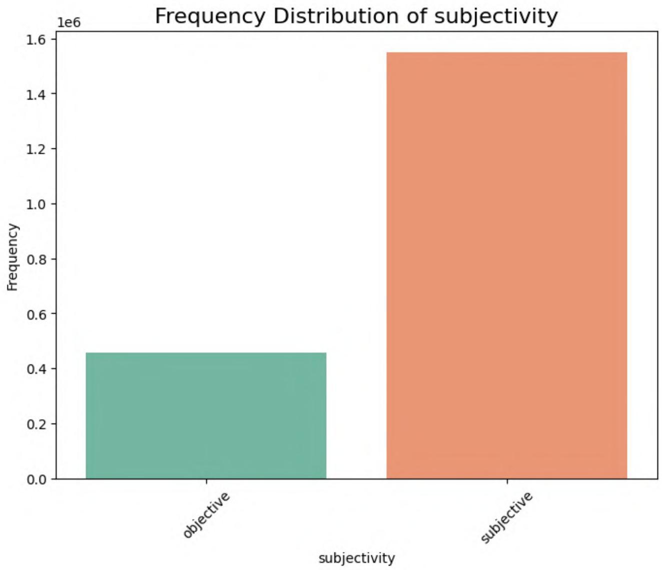

Fig. 11: Frequency plot for subjectivity column

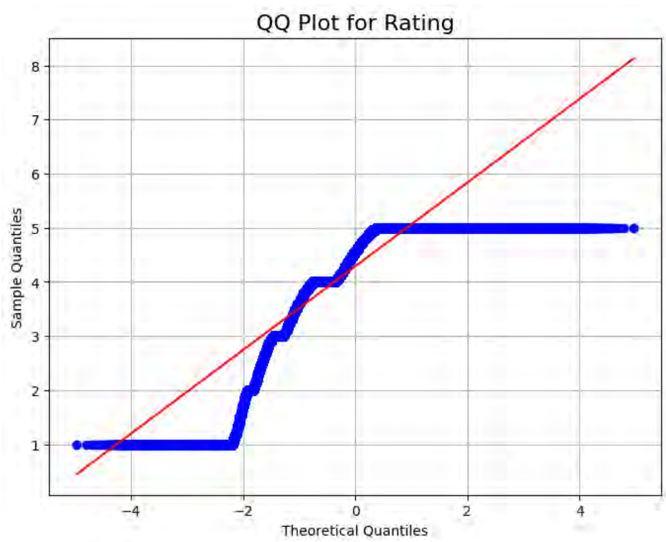

Fig. 12: QQ plot for Rating column

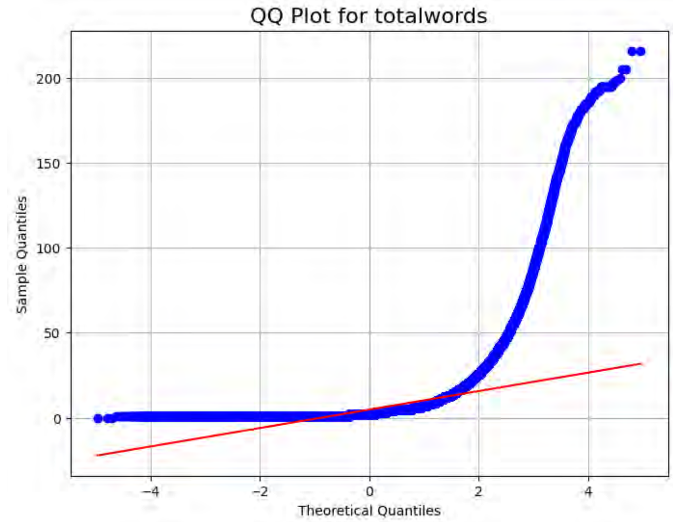

Fig. 13: QQ plot for totalwords column

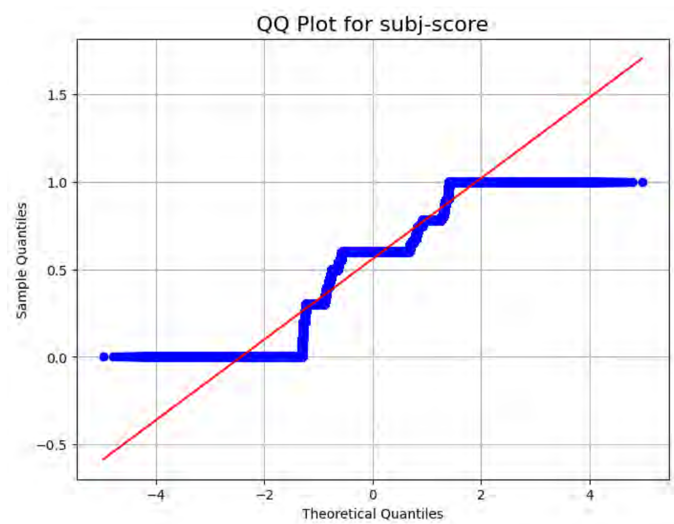

Fig. 14: QQ plot for Subjectivity-score column

is concluded that the means of the groups are significantly different.

5) Data Distribution Analysis: The Quantile-Quantile (QQ) plot is analyzed to depict the current TPE dataset data distribution. A Quantile-Quantile (QQ) plot is a graphical tool used to assess if a dataset follows a specific theoretical distribution, such as the normal distribution. The QQ plot compares the dataset's quantiles against the theoretical distribution's quantiles. If the data follows the theoretical distribution, the points on the QQ plot will lie close to a straight line. Deviations from the straight line indicate departures from the theoretical distribution. Fig. 12, Fig. 13, and Fig. 14 show the QQ plots for the Ratting, totalwords and subjectivity score numerical columns.

The figures show that the plots deviate from a straight line, suggesting that the data do not follow a normal distribution. As the data did not follow a complete normal distribution, non-parametric tests, namely Mann-Whitney U and Kruskal-Wallis tests, are executed. The results from the test are discussed as follows.

Mann-Whitney U test results

Positive vs. Neutral

Mann-Whitney U statistic: 248,843,245,695.0

P-value: 0.0

Positive vs. Negative

Mann-Whitney U statistic: 243,157,072,917.0

P-value: 0.0

The Mann-Whitney U test, or the Wilcoxon rank-sum test, compares two independent groups. In this case, we have compared the "Rating" variable for the "positive" sentiment group with the "neutral" sentiment group and with the "negative" sentiment group. The test yields two important results: the Mann-Whitney U statistic and the p-value. The Mann-Whitney U statistic represents the rank-sum of one group (positive) relative to the other (neutral or negative) group. It indicates the sum of ranks assigned to the observations in the positive group. The smaller the U statistic, the more likely the two groups differ significantly. The p-value measures the evidence against the null hypothesis (the assumption that there is no difference between the groups). A p-value of 0.0 indicates that there is extremely strong evidence to reject the null hypothesis and conclude that there is a significant difference between the groups' ratings. In other words, the "positive" sentiment group has significantly different ratings compared to both the "neutral" and "negative" sentiment groups.

Kruskal-Wallis test results

Kruskal-Wallis H statistic: 777,617.6570663975

P-value: 0.0

The Kruskal-Wallis test is a non-parametric alternative to the one-way ANOVA used to compare more than two independent groups. In this case, we are comparing the "Rating" variable for all three sentiment groups: "positive", "neutral," and "negative". The test provides the Kruskal-Wallis H statistic and the p-value. The Kruskal-Wallis H statistic measures the degree of variation between the groups. The larger the H statistic, the more evidence suggests that at least one group differs significantly from the others regarding the "Rating" variable. The p-value of 0.0 indicates that there is strong evidence to reject the null hypothesis and conclude that there is a significant difference between the groups' ratings. Therefore, we can infer that the sentiment groups ("positive," "neutral," and "negative") have significantly different ratings based on the Kruskal-Wallis test.

G. Dataset labeling and Inter Annotator Agreement

Dataset labeling is a crucial step in any machine learning task, especially in sentiment analysis where the model's accuracy heavily depends on the quality of the labeled data. In this study, we labeled the dataset using a team of undergraduate students trained to identify the sentiment of different aspects of teacher performance. The labelling was done on a scale of three classes, namely positive, negative, and neutral, to provide a comprehensive understanding of the sentiment conveyed in the dataset.

To ensure the quality of the labeled data, we calculated the Fleiss' kappa score, a measure of inter-rater agreement among multiple annotators. Fleiss' kappa is a widely used statistical measure to evaluate the agreement between multiple raters and has been extensively used in sentiment analysis studies. In our study, we achieved a Fleiss' kappa score of over 94%, which indicates a high level of agreement among the annotators.

Table VIII shows the Fleiss' kappa score obtained for our labeled dataset. The table presents the agreement score for each aspect of teacher performance and the overall agreement score for the entire dataset. As we can see, the Fleiss' kappa score for all aspects is above 0.9, indicating almost perfect agreement among the annotators. The overall Fleiss' kappa score for the dataset is 0.947, considered an excellent level of agreement.

The high level of agreement among the annotators in labeling the dataset is a testament to the quality of the labeled data used in this study. The labeled dataset provides a reliable and accurate data source for training the ABSA LSTM model, and we believe that the model's performance is a direct result of the quality of the labeled data.

TABLE VIII: Fleiss' Kappa Score for Labeled Dataset

Aspect	Positive	Negative	Neutral	Agreement
Knowledge	0.934	0.932	0.944	0.937
Clarity	0.947	0.936	0.940	0.941
Approachability	0.951	0.946	0.942	0.946
Fairness	0.943	0.942	0.948	0.944
Overall	0.950	0.948	0.951	0.947

H. Data Availability

The dataset is available upon request to the corresponding author or can be obtained directly from Mendeley Data [30]. We believe this dataset will be valuable for researchers, educational institutions, and policymakers using sentiment analysis for teacher performance evaluation. The dataset provides a more reliable and objective approach to evaluating teacher performance, leading to more informed decisions and improving the quality of education. Furthermore, we hope this dataset will encourage further research in the field of sentiment analysis for teacher performance evaluation and contribute to developing more accurate and effective evaluation methods.

I. Experimental Setup

The experiments were conducted on a system with a Xeon processor, 500GB SSD and 512GB of RAM, running Ubuntu operating system. The hardware specifications of the system used for experiments are given in Table 1. The software requirements for running the experiments were Python 3.7, Keras 2.4.3, TensorFlow 2.3.1 and Pandas 1.0.3. The dataset was stored on Google Drive and accessed using the PyDrive library. The experiments were conducted in a Jupyter Notebook environment. Table IX shows the hardware configuration utilized in this study.

TABLE IX: Hardware Specifications

Processor	Intel Xeon
CPU Cores	24
Clock Speed	2.5 GHz
RAM	512 GB
Storage	500 GB SSD

The hardware used for the experiments provided sufficient computational power to run the necessary python scripts efficiently. The system had enough RAM to handle large datasets and the SSD provided fast read and write speeds, which helped load the dataset quickly. The processor with 16 cores and a clock speed of 2.5 GHz allowed the model to train quickly, reducing the overall experiment time. The Ubuntu operating system was chosen for its stability and ease of use. The software requirements for running the experiments were all open-source and readily available for download, which made it easy to set up the experimental environment. The Jupyter Notebook environment provided an interactive and user-friendly interface for running the experiments and analyzing the results.

IV. Experimental Results Analysis

The dataset is evaluated with different baseline machine learning models to check how different baseline models perform with respect to accuracy and F1 while detecting fraud reviews. Fifteen models from scikit learn library has been tested with the dataset. The models are, SVC, Random Forest Classifier, Gaussian NB, Bernoulli NB, SGD Classifier, Perceptron, Ridge Classifier CV, Ridge Classifier, Linear SVC, Calibrated Classifier CV, Logistic Regression, Linear Discriminant Analysis, Passive Aggressive Classifier, Quadratic Discriminant Analysis, and AdaBoost Classifier. The comparative results are presented in Table X.

The dataset is assessed using various baseline machine learning models, aiming to analyze the performance of these models concerning the accuracy and F1-score in the context of fraud review detection. A comprehensive set of fifteen distinct models sourced from the scikit-learn library is employed for experimentation on the given dataset. The roster of models encompasses Support Vector Classifier (SVC), Random Forest Classifier, Gaussian Naive Bayes (Gaussian NB), Bernoulli Naive Bayes (Bernoulli NB), Stochastic Gradient Descent (SGD) Classifier, Perceptron, Ridge Classifier with Cross-Validation (Ridge Classifier CV), Ridge Classifier, Linear Support Vector Classifier (Linear SVC), Calibrated Classifier with Cross-Validation (Calibrated Classifier CV), Logistic Regression, Linear Discriminant Analysis, Passive Aggressive Classifier, Quadratic Discriminant Analysis, and AdaBoost Classifier.

Through a comprehensive analysis, the performance of these models is evaluated and compared based on their efficacy in detecting fraudulent reviews. The evaluation metrics employed for comparison encompass accuracy and F1-score. The outcomes of this comparative analysis are succinctly presented in Table X, providing a consolidated perspective on the relative capabilities of the diverse baseline machine learning models in the specific context of fraud review detection.

TABLE X: Comparative results of different baseline models

Model Name	Accuracy	Balanced Accuracy	F1	Time Taken in seconds
SVC	0.9	0.89	0.89	3778.96
Random Forest Classifier	0.96	0.91	0.91	51.57
Gaussian NB	0.95	0.97	0.95	6.68
Bernoulli NB	0.94	0.95	0.94	8.65
SGD Classifier	0.96	0.94	0.96	6.39
Perceptron	0.95	0.91	0.95	6.26
Ridge Classifier CV	0.95	0.9	0.95	9.13
Ridge Classifier	0.95	0.9	0.95	8.32
Linear SVC	0.95	0.9	0.95	248.71
Calibrated Classifier CV	0.95	0.9	0.95	28.87
Logistic Regression	0.94	0.9	0.94	8.72
Linear Discriminant Analysis	0.94	0.89	0.94	9.27
Passive Aggressive Classifier	0.93	0.88	0.93	6.67
Quadratic Discriminant Analysis	0.61	0.76	0.65	7.48
AdaBoost Classifier	0.86	0.68	0.84	118.02

The table provides insights into the accuracy, balanced accuracy, F1-score, and computational time each model takes. These metrics serve as crucial indicators to evaluate the effectiveness of each model in the context of fraud detection.

The models' accuracy scores range from 0.61 to 0.96, showcasing a notable variation in their predictive capabilities. Among the models, the Random Forest Classifier stands out with a commendable accuracy of 0.96, indicating its proficiency in correctly classifying fraudulent reviews. The Gaussian NB and SGD Classifiers follow closely, achieving an accuracy of 0.95. These high accuracy scores underscore the models' adeptness in distinguishing between genuine and fraudulent reviews, which is imperative in maintaining the credibility of online platforms.

The balanced accuracy metric is considered to assess the models' performance further. This metric accounts for any class imbalances within the dataset and provides a more comprehensive understanding of a model's ability to generalize across classes. Interestingly, while the Random Forest Classifier continues to excel with a balanced accuracy of 0.91, the Gaussian NB outperforms other models with an impressive balanced accuracy of 0.97. These results reaffirm the robustness of Gaussian NB in mitigating class imbalances and making accurate predictions.

The F1 scores, a harmonic mean of precision and recall, also offer valuable insights into model performance. Models such as Random Forest Classifier, Gaussian NB, SGD Classifier, Perceptron, Ridge Classifier CV, Ridge Classifier, Linear SVC, and Calibrated Classifier CV demonstrate consistent F1-scores of 0.91 or 0.95, indicating their balanced precision and recall in detecting fraudulent reviews.

Computational time is a significant consideration in real-world applications, as it impacts the efficiency of model deployment. The models exhibit varying time requirements, ranging from a few seconds to several minutes. The Quadratic Discriminant Analysis and AdaBoost Classifier demand relatively longer computation times, with 7.48 and 118.02 seconds, respectively. Conversely, models like Gaussian NB, Bernoulli NB, and SGD Classifier exhibit low time requirements, making them more suitable for applications requiring swift fraud detection.

The proposed dataset, meticulously compiled and evaluated with the baseline ML models stands as a testament to its comprehensiveness and profound significance within the fraud detection domain, specifically concerning the evaluation of teachers' performance. The imperative necessity for such a dataset becomes conspicuously apparent within the dynamically evolving educational milieu, where the accurate and equitable appraisal of teachers' instructional efficacy holds intrinsic value. The dataset squarely addresses this exigent requirement and transcends prevailing benchmarks in the realm of fraud detection germane to teacher evaluation. The dataset's adaptability, underscored by the superlative performance exhibited by sundry baseline models, lucidly underscores its prospective potential to catalyze a paradigm shift within prevailing fraud detection paradigms, notably within the intricate milieu of pedagogic performance assessment, thereby eliciting a transformative effect upon the landscape of educational caliber assurance.

V. Conclusion

This article presented a novel dataset for aspect-based sentiment analysis for teacher performance evaluation and the total process of creating the dataset. Our study highlights the importance of good datasets for aspect-based sentiment analysis and the potential of this approach for improving teaching effectiveness and student outcomes.

This study contributes to the growing body of research on sentiment analysis and its applications in education. We hope our findings inspire further research and innovation and improve teaching effectiveness and student outcomes.

References

[1] V. Kuleto, M. Ilić, M. Dumangiu, M. Ranković, O. M. D. Martins, D. Păun, and L. Mihoreanu, "Exploring opportunities and challenges of artificial intelligence and machine learning in higher education institutions," *Sustainability*, vol. 13, no. 18, 2021. [Online]. Available: https://www.mdpi.com/2071-1050/13/18/10424

[2] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowledge-Based Systems*, vol. 226, p. 107134, Aug. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S095070512100397X

[3] M. S. U. Miah, J. Sulaiman, T. B. Sarwar, A. Naseer, F. Ashraf, K. Z. Zamli, and R. Jose, "Sentence boundary extraction from scientific literature of electric double layer capacitor domain: tools and techniques," *Applied Sciences*, vol. 12, no. 3, p. 1352, 2022.

[4] M. Hoang, O. A. Bihorac, and J. Rouces, "Aspect-based sentiment analysis using BERT," in *Proceedings of the 22nd Nordic Conference on Computational Linguistics*. Turku, Finland: Linköping University Electronic Press, Sept.–Oct. 2019, pp. 187–196. [Online]. Available: https://aclanthology.org/W19-6120

[5] H. Peng, L. Xu, L. Bing, F. Huang, W. Lu, and L. Si, "Knowing What, How and Why: A Near Complete Solution for Aspect-Based Sentiment Analysis," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, pp. 8600–8607, Apr. 2020. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/6383

[6] N. C. Dang, M. N. Moreno-García, and F. De La Prieta, "Sentiment Analysis Based on Deep Learning: A Comparative Study," *Electronics*, vol. 9, no. 3, p. 483, Mar. 2020. [Online]. Available: https://www.mdpi.com/2079-9292/9/3/483

[7] M. Saef Ullah Miah and J. Sulaiman, "Material named entity recognition (mner) for knowledge-driven materials using deep learning approach," in *Proceedings of the Fourth International Conference on Trends in Computational and Cognitive Engineering: TCCE 2022*. Springer, 2023, pp. 199–208.

[8] M. S. Ullah Miah, J. Sulaiman, T. B. Sarwar, S. S. Islam, M. Rahman, and M. S. Haque, "Medical named entity recognition (medner): A deep learning model for recognizing medical entities (drug, disease) from scientific texts," in *IEEE EUROCON 2023 - 20th International Conference on Smart Technologies*, 2023, pp. 158–162.

[9] M. S. U. Miah, J. Sulaiman, T. B. Sarwar, I. U. Ferdous, S. S. Islam, and M. S. Haque, "Target and precursor named entities recognition from scientific texts of high-temperature steel using deep neural network," in *Database and Expert Systems Applications*, C. Strauss, T. Amagasa, G. Kotsis, A. M. Tjoa, and I. Khalil, Eds. Cham: Springer Nature Switzerland, 2023, pp. 203–208.

[10] L. Yang, Y. Li, J. Wang, and R. S. Sherratt, "Sentiment analysis for e-commerce product reviews in chinese based on sentiment lexicon and deep learning," *IEEE Access*, vol. 8, pp. 23 522–23 530, 2020.

[11] L. Nemes and A. Kiss, "Social media sentiment analysis based on covid-19," *Journal of Information and Telecommunication*, vol. 5, no. 1, pp. 1–15, 2021. [Online]. Available: https://doi.org/10.1080/24751839.2020.1790793

[12] E. Park, J. Kang, D. Choi, and J. Han, "Understanding customers' hotel revisiting behaviour: a sentiment analysis of online feedback reviews," *Current Issues in Tourism*, vol. 23, no. 5, pp. 605–611, 2020. [Online]. Available: https://doi.org/10.1080/13683500.2018.1549025

[13] S. K. Carpenter, A. E. Witherby, and S. K. Tauber, "On students' (mis)judgments of learning and teaching effectiveness," *Journal of Applied Research in Memory and Cognition*, vol. 9, no. 2, pp. 137–151, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2211368120300024

[14] S. L. Wallace, A. K. Lewis, and M. D. Allen, "The state of the literature on student evaluations of teaching and an exploratory analysis of written comments: Who benefits most?" *College Teaching*, vol. 67, no. 1, pp. 1–14, 2019. [Online]. Available: https://doi.org/10.1080/87567555.2018.1483317

[15] A. ONAN, "Sentiment analysis on massive open online course evaluations: A text mining and deep learning approach," *Computer Applications in Engineering Education*, vol. 29, no. 3, pp. 572–589, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cae.22253

[16] Z. Drus and H. Khalid, "Sentiment analysis in social media and its application: Systematic literature review," *Procedia Computer Science*, vol. 161, pp. 707–714, 2019, the Fifth Information Systems International Conference, 23-24 July 2019, Surabaya, Indonesia. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S187705091931885X

[17] A. U. Rehman, A. K. Malik, B. Raza, and W. Ali, "A Hybrid CNN-LSTM Model for Improving Accuracy of Movie Reviews Sentiment Analysis," *Multimedia Tools and Applications*, vol. 78, no. 18, pp. 26 597–26 613, Sept. 2019. [Online]. Available: http://link.springer.com/10.1007/s11042-019-07788-7

[18] H. H. Do, P. Prasad, A. Maag, and A. Alsadoon, "Deep Learning for Aspect-Based Sentiment Analysis: A Comparative Review," *Expert Systems with Applications*, vol. 118, pp. 272–299, Mar. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0957417418306456

[19] P. K. Jain, R. Pamula, and G. Srivastava, "A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews," *Computer Science Review*, vol. 41, p. 100413, Aug. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1574013721000538

[20] R. Dotan and S. Milli, "Value-laden disciplinary shifts in machine learning," *arXiv preprint arXiv:1912.01172*, 2019.

[21] E. Denton, A. Hanna, R. Amironesei, A. Smart, H. Nicole, and M. K. Scheuerman, "Bringing the people back in: Contesting benchmark machine learning datasets," *arXiv preprint arXiv:2007.07399*, 2020.

[22] A. Bhowmik, M. S. U. Miah, *et al.*, "Iot (internet of things)-based smart garbage management system," *AIUB Journal of Science and Engineering (AJSE)*, vol. 19, no. 1, pp. 33–40, 2020.

[23] T. Simonite, "Google's ai guru wants computers to think more like brains. 2018," *URL https://www. wired. com/story/googles-ai-guru-computers-think-more-like-brains*.

[24] M. T. Hossain, A. Hossain, S. M. Meem, M. F. Monir, M. S. Ullah Miah, and T. Bin Sarwar, "Impact of covid-19 lockdowns on air quality in bangladesh: Analysis and aqi forecasting with support vector regression," in *2023 4th International Conference for Emerging Technology (INCET)*, May 2023, pp. 1–6.

[25] B. Hutchinson, A. Smart, A. Hanna, E. Denton, C. Greer, O. Kjartansson, P. Barnes, and M. Mitchell, "Towards accountability for machine learning datasets: Practices from software engineering and infrastructure," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 560–575.

[26] Y. Roh, G. Heo, and S. E. Whang, "A survey on data collection for machine learning: a big data-ai integration perspective," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 4, pp. 1328–1347, 2019.

[27] J. McArthur, N. Shahbazi, R. Fok, C. Raghubar, B. Bortoluzzi, and A. An, "Machine learning and bim visualization for maintenance issue classification and enhanced data collection," *Advanced Engineering Informatics*, vol. 38, pp. 101–112, 2018.

[28] F. Barbieri, J. Camacho-Collados, L. Neves, and L. Espinosa-Anke, "Tweeteval: Unified benchmark and comparative evaluation for tweet classification," *arXiv preprint arXiv:2010.12421*, 2020.

[29] S. Loria, "Textblob: Simplified text processing," https://textblob.readthedocs.io/en/dev/, 2018, accessed on 20 April 2023.

[30] A. Bhowmik, N. M. Noor, M. S. U. Miah, M.-U. Haque, and D. Karmaker, "A novel dataset for aspect-based sentiment analysis for teacher performance evaluation," 2023.

Abhijit Bhowmik completed his B.Sc. in Computer Science & Engineering in 2009 and M.Sc. in Computer Science in 2011 from the American International University– Bangladesh (AIUB). He is pursuing his PhD from University Malaysia Pahang Al-Sultan Abdullah in NLP and Machine Learning. He is an Associate Professor and Special Assistant Office of Student Affairs (OSA) in the Department of Computer Science, AIUB. His research interests include NLP, Machine Learning, software engineering, mobile & multimedia communication, and data mining. Mr. Bhowmik can be contacted at ovi775@gmail.com.

Noorhuzaimi Mohd Noor has been in the academic, research & consultancy field since 2003. She is Head of the Program (Entrepreneurship) at the Centre of Creative Entrepreneur Development and Senior Lecturer at The Universiti Malaysia Pahang Al-Sultan Abdullah, Malaysia. She received her B.Sc. in Computer Science from the Universiti Putra Malaysia, Malaysia, in 1999, followed by a Master's in Science from the same university in 2003. She received her PhD in Computer Sciences from Universiti Kebangsaan Malaysia, Malaysia 2016. She is the author of more than 20 research articles. Her research interests include natural language processing, expert system, and computer security. She is also a Reviewer for the Journal of Information and Communication Technology (JICT) and editor for the International Journal of Software Engineering and Computer Systems (IJSECS). Dr. Noorhuzaimi is also a certified Professional Technologist from the Malaysia Board of Technologists (MBOT), where he is actively involved as Assessor Panel for Technology and Technical Academic Programs Accreditation.

M. Saef Ullah Miah is working as an assistant professor in the Department of Computer Science at American International University-Bangladesh (AIUB). He is currently engaged in research and teaching activities and has practical experience in software development and project management. He obtained his PhD from Universiti Malaysia Pahang and earned his Master of Science and Bachelor of Science degrees from AIUB. In addition to his professional activities, he is passionate about working on various open-source projects. His main research interests are data and text mining, natural language processing, machine learning, material informatics and blockchain applications.

Md. Mazid-Ul-Haque is currently working as a Lecturer in the Department of Computer Science at American International University-Bangladesh (AIUB). He has completed both his Master of Science in Computer Science degree and Bachelor of Science in Computer Science and Engineering degree from AIUB with the highest honor and academic awards. He did his HSC at Notre Dame College, Dhaka. He has a strong passion and dedication for teaching and research work. His research interests include but are not limited to Network, Wireless Communication, SDLC.

Debajyoti Karmaker is working as an Associate professor in the Department of computer science at American International University-Bangladesh. He worked as Postdoctoral Research Fellow at Australian National University (ANU), and Stanford University. Before joining ANU, he completed PhD from The University of Queensland (UQ). His research interests are Deep Learning, Computer Vision, & Machine Learning. He is particularly interested in image classification, object detection, segmentation, bio-inspired collision avoidance strategies, and Robust Decision-making and Learning. Before starting his PhD, he worked as a Lecturer at the American International University-Bangladesh (AIUB) - in the Department of Computer Science. He also worked as a software engineer at Infra Blue Technology (IBT Games).

APPENDIX

Student Comments	Rating	Course ID	Offered Course ID	Course Name	Section	Semester
gjfgjgfjkjk	4.58			COMMUNICAT	COMMUNICATI	2003-2004, Summer
good	4.96			PRICING STRAT	PRICING STRATE	2003-2004, Summer
good	5			STRATEGIC MA	STRATEGIC MAR	2003-2004, Summer
gjfgjgf	4.71			DATA COMMU	DATA COMMUN	2003-2004, Summer
gjfgjgf	4.71			DATA COMMU	DATA COMMUN	2003-2004, Summer
TEACHER	4.25			STRENGTH OF	STRENGTH OF M	2003-2004, Summer
gjf	4.58			ARTIFICIAL INT	ARTIFICIAL INTE	2003-2004, Summer
friendly teacher but not enough ability to encoura	4.38			SYSTEM PROGR	SYSTEM PROGR	2003-2004, Summer
TEACHER	4.92			SYSTEM PROGR	SYSTEM PROGR	2003-2004, Summer
hhhdfhg	4.67			AUTOMATA TH	THEORY OF COM	2003-2004, Summer
hhhdfhg	4.67			THEORY OF CO	THEORY OF COM	2003-2004, Summer
he is a good techer.	5			MANAGEMENT	MANAGEMENT	2003-2004, Summer
he is a good techer.	5			SOFTWARE EN	SOFTWARE ENG	2003-2004, Summer
he is agood techer.	4.81			SOFTWARE EN	SOFTWARE ENG	2003-2004, Summer
TEACHER	4.75			OBJECT ORIENT	PROGRAMMING	2003-2004, Summer
TEACHER	4.75			PROGRAMMIN	PROGRAMMING	2003-2004, Summer
TEACHER	4.7			DIGITAL ELECTF	LOGIC DESIGN 2	2003-2004, Summer
TEACHER	4.7			LOGIC DESIGN	LOGIC DESIGN 2	2003-2004, Summer
he is a good teacher	5			PRINCIPLES OF	PRINCIPLES OF	2003-2004, Summer

Fig. 15: Structure of the initial raw data collected from the university system

	StudentComments	Rating	totalwords	Sentiment	sent_pretrain	subjectivity	subj-score	isSame
0	good	4.96	1	positive	positive	subjective	0.6	TRUE
1	good	5	1	positive	positive	subjective	0.6	TRUE
2	teacher	4.25	1	positive	neutral	objective	0	fake
3	friendly teacher but not	4.38	10	positive	neutral	subjective	0.5	fake
4	teacher	4.92	1	positive	neutral	objective	0	fake
5	he is a good techer.	5	5	positive	positive	subjective	0.6	TRUE
6	he is a good techer.	5	5	positive	positive	subjective	0.6	TRUE
7	he is agood techer.	4.81	4	positive	neutral	objective	0	fake
8	teacher	4.75	1	positive	neutral	objective	0	fake
9	teacher	4.75	1	positive	neutral	objective	0	fake
10	teacher	4.7	1	positive	neutral	objective	0	fake
11	teacher	4.7	1	positive	neutral	objective	0	fake
12	he is a good teacher	5	5	positive	positive	subjective	0.6	TRUE
13	above all our teacher is a	4.75	8	positive	positive	objective	0.1	TRUE
14	excellent teacher. great	5	4	positive	positive	subjective	0.875	TRUE
15	have excellent attitude,	3.88	12	positive	positive	subjective	0.875	TRUE
16	he is very good for our st	4.04	11	positive	positive	subjective	0.55666667	TRUE
17	he is a good teacher.	4.54	5	positive	positive	subjective	0.6	TRUE
18	he is a good teacher	3.63	5	positive	positive	subjective	0.6	TRUE
19	he is a good teacher and	4.49	15	positive	positive	objective	0.4	TRUE
20	he is a very good teache	5	6	positive	positive	subjective	0.78	TRUE
21	he is a very good teache	5	7	positive	positive	subjective	0.78	TRUE
22	he is efficient and bette	4.5	15	positive	positive	objective	0.4	TRUE
23	he is great and honest.	3	5	neutral	positive	subjective	0.825	fake
24	he is the best.	4.83	4	positive	positive	objective	0.3	TRUE
25	he is the well known tec	4.42	11	positive	positive	objective	0.2	TRUE
26	he is very good teacher.	4.42	11	positive	positive	objective	0.44	TRUE
27	i want him as a faculty in	4.17	9	positive	neutral	objective	0	fake
28	mr. mohiuddin is an exc	5	6	positive	positive	objective	0	TRUE

Fig. 16: Structure of the final curated TPE datset