Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 2 (2025)

Hybrid Lexicon and Transformer-Based Sentiment Analysis of Student Feedback for Faculty Evaluation: A Speech-to-Text Approach

Helaria Maria[1], R Subhashni[2]

[1]Research Scholar, Department of Computer Applications St.Peter's Institute of Higher Education & Research Chennai. Email: helariax@gmail.com

[2]Department of Computer Applications St.Peter's Institute of Higher Education & Research Chennai. Email: subhashniraj2018@gmail.com

Article History:

Received: *13-07-2024*

Revised: *28-08-2024*

Accepted: *26-09-2024*

Abstract:

The conventional methods for assessing the performance of educators in academic settings have often been hindered by outdated feedback systems, leading to assessments that can be either skewed or in- complete. This study unveils an innovative, dual-approach methodology that fuses lexicon-driven sentiment analysis with cutting-edge Transformer architectures, specifically focusing on BERT (Bidirectional Encoder Representations from Transformers). This approach provides a more sophisticated, real-time assessment of educators based on student evaluations. One of the standout features of this study is the integration of Speech-to-Text technologies, which allows for the immediate transformation of verbal feedback into text that can be analyzed. The approach makes use of a specialized Educational Sentiment Lexicon for preliminary sentiment evaluation, which subsequently refines the performance of a pre-existing BERT model. This dual-model is proficient in scrutinizing both text-based and verbal feedback—the latter being converted into text through advanced Speech-to-Text techniques. Our results indicate that this approach significantly outperforms existing lexicon-based and machine learning methods in terms of accuracy and comprehensiveness. By providing a real-time, versatile sentiment analysis tool tailored for educational settings, this research marks a paradigm shift in the scope and quality of faculty evaluations, thereby contributing substantially to the field of educational technology and sentiment analysis.

Keywords: Educational Sentiment Analysis, Faculty Performance Evaluation, Hybrid Model, Lexicon- Based Model, BERT-Based Model, Real-Time Evaluation, Student Feedback, Real-Time Sentiment Analysis Tool.

1 Introduction

In the ever-changing landscape of education, the evaluation of faculty performance has always been a cornerstone for institutional improvement. Traditionally, these evaluations have been conducted through written surveys, peer reviews, and administrative assessments (1). However, these methods often suffer from various limitations such as low response rates, potential biases, and the inability to capture the nuanced sentiments of students. The emergence of machine learning and natural language processing (NLP) technologies has paved the way for more dynamic, instantaneous, and impartial evaluations. Sentiment analysis, a specialized branch of NLP, offers a particularly promising path. This technique specializes in deciphering and categorizing emotional undertones in textual content (2). Its application in educational contexts, especially for appraising faculty based on student feedback, is a budding yet under-researched domain

A. The Intricacies of Student Feedback and the Need for Sentiment Analysis

Student evaluations offer a treasure trove of data that can shed light on the effectiveness of educators. Traditional approaches to gathering and interpreting this feedback, however, often miss the mark. Paper-based feedback forms are labor-intensive to scrutinize and are prone to various biases (3). For instance, students might not be forthright or articulate their sentiments am-biguously, leading to misunderstandings. Sentiment analysis can fill this void. Utilizing machine learning techniques on textual data, sentiment analysis can offer a more intricate comprehension of student evaluations, thereby enhancing the reliability of faculty assessments.

B. Lexicon-Based Sentiment Analysis: A Focused Approach

A particularly potent strategy in sentiment analysis is the lexicon-centric method. This ap- proach employs a predetermined list of words, each tagged with a sentiment score to denote its emotional tone—either positive, negative, or neutral (4). This technique is especially relevant in educational contexts where feedback can be highly specialized. However, most available lexiconsare universal and may not effectively capture the subtleties of educational evaluations. Thisstudy seeks to fill this void by crafting a specialized Educational Sentiment Lexicon designedto scrutinize student evaluations of faculty performance.

C. Bridging the Gap: Spoken Feedback and Real-Time Analysis

Another innovative facet of this study is its emphasis on oral student evaluations. Whilewritten evaluations have been the subject of extensive research, oral feedback remains largely unexplored. Oral evaluations offer the benefit of being more immediate and, often, more candid. However, they also introduce unique challenges, such as the necessity for precise speech-to-text translation and the intricacy of dissecting spoken language, which may lack the grammaticalrigor of written text. This study aims to overcome these challenges by utilizing state-of-the-art speech-to-text algorithms to transcribe oral feedback into textual data. This data is then analyzed using our specialized Educational Sentiment Lexicon and sophisticated machine learning tech- niques. The ultimate objective is to establish a real-time sentiment analysis framework that can offer immediate insights into educator effectiveness, thereby enabling prompt interventions and enhancements.

2. LITERATURE REVIEW

The burgeoning field of sentiment analysis in educational contexts has seen a plethora of research endeavors aimed at leveraging machine learning and natural language processing tech- niques to evaluate faculty performance based on student feedback (6). These studies employ a myriad of methodologies and technologies, each with its own set of merits and limitations. This Literature Review section aims to provide an exhaustive critique of the existing body of work, thereby laying the groundwork for our own research, which seeks to address the gaps identified herein.

Ochilbek Rakhmanov et al (7) explained the use of machine learning algorithms, ranging from traditional methods like Random Forest and Support Vector Machines (SVM) to more advanced techniques like Long Short-Term Memory (LSTM), bi-directional LSTM, BERT, and RoBERTa. However, a glaring limitation of this study is the absence of specific numerical results. This omission hampers the ability to gauge the effectiveness of the algorithms employed, thereby

leaving a gap in the literature that future research could aim to fill. The lack of quantitative data also raises questions about the replicability of the study, a cornerstone of scientific research.

Venkateswarlu Bonta et al (8) employed Natural Language Toolkit (NLTK), TextBlob, and VADER to conduct sentiment analysis and conclude that VADER outperforms TextBlob. While the study provides valuable insights into the performance of different lexicons, it does not delve into the limitations or potential biases inherent in these lexicons. This omission is particularly glaring given that educational feedback can be highly context-specific and laden with jargon and idioms that generic lexicons may not adequately capture.

Irum Sindhu et al (9) delves into the realm of aspect-based sentiment analysis, employinga two-layered Long Short-Term Memory (LSTM) network. The study reports an accuracy of91% for aspect extraction and 93% for sentiment polarity. While these results are promising, the paper does not address the computational cost associated with deep learning models. This is a critical oversight, as the feasibility of implementing such models in real-time educational settingsremains an open question. Moreover, the study does not explore alternative machine learning algorithms that might offer comparable performance with lower computational overhead.

Ganpat Singh Chauhan et al (10) utilized Naive Bayes and Part-of-Speech (POS) taggers to conduct its analysis. It reports an F-measure of 0.80 for teaching aspects and 0.81 for course aspects. While the paper provides valuable insights into the utility of POS tagging in sentiment analysis, it is limited by its narrow focus on only one machine learning algorithm. This limitationrestricts the generalizability of the findings and leaves room for future research to explore the efficacy of other algorithms in similar settings.

Aytug ONAN et al (11) employed a range of machine learning algorithms including Con- volutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Bi-directional RNN, Gated Recurrent Units (GRU), and LSTM. It reports the highest accuracy of 95.80% usingLSTM in conjunction with GloVe embeddings. However, the paper falls short in discussingthe computational resources required for implementing these deep learning models. This is a significant gap, given that educational institutions may not always have access to the high- performance computing resources required for such models.

Rdouan Faizi et al (12) employed VADER and an educational sentiment lexicon to improve the accuracy to 86.45%. While the study provides valuable insights into the utility of domain- specific lexicons, it lacks a comparative analysis with other machine learning algorithms. This omission limits the study's contribution to the broader field of sentiment analysis in educational

settings.

Muhammad Umair and team (13) employed a range of algorithms such as Support Vector Machine, Na¨ıve Bayes, TextBlob, VADER, and NLTK, and they reported an average accuracy rate of 85.62% when using SVM with K-fold cross-validation. Although the study offers an extensive survey of multiple algorithms, it neglects to address the limitations or potential biases that may arise when these algorithms are used in the context of educational feedback. This omission is noteworthy, as generic algorithms may not fully grasp the intricacies and subtletiesof student feedback, which could impact the precision of sentiment analysis.

Zenun Kastrati and collaborators (14) made use of Convolutional Neural Networks (CNN)and Long Short-Term Memory (LSTM) networks. However, their study is devoid of quantitativemetrics. They characterize their model's performance as "inspiring," but the lack of numerical data hampers the ability to assess the efficacy of their methodology. This creates a void in research that could be filled by future studies offering empirical data.

In a similar vein, Pansy Nandwani and associates (15) present an overview of multipletechniques but do not go the extra mile to offer a side-by-side comparison. While their paper serves as an overview, it lacks the critical analysis that could be beneficial for researchersseeking empirical evidence on the effectiveness of different algorithms. This limitation restricts the paper's contribution to the field and calls for more comprehensive comparative studies.

Another work by Zenun Kastrati et al (16) reviews various techniques and provides an analysisof educational entities, research trends, and methodologies. However, the paper lacks specific recommendations for future research, leaving a significant gap in the literature. The absence of actionable insights limits the paper's utility for researchers aiming to build upon existing work.

A. Limitations in Existing Literature

A common thread running through the reviewed literature is the lack of comprehensive com- parative studies that evaluate the performance of different algorithms and lexicons in educationalsettings. This is a significant gap, as comparative analyses are crucial for understanding the rela- tive strengths and weaknesses of various approaches. Additionally, most papers do not discuss thecomputational costs associated with their proposed models. This is a critical oversight, especiallyfor real-time applications where computational efficiency is paramount. Another recurring issue is the absence of specific numerical results in some studies, which hampers the ability to assess their efficacy.

3. METHODOLOGY

In our methodology, we utilize TensorFlow to implement a meta-classifier that combines Lexicon-Based and BERT-Based models through a Gradient Boosting Machine (GBM), opti- mized for imbalanced datasets. WordPiece tokenization, special tokens, and fine-tuning of the BERT-Base model are executed within the TensorFlow framework. We apply rigorous 5-fold cross-validation for model validation, calculated through TensorFlow-enabled custom metrics likeaccuracy, F1-score, and AUC-ROC. This TensorFlow-backed methodology sets new standardsin educational sentiment analysis and is quantitatively described for clarity.

A. Data Collection

The efficacy of any sentiment analysis research is intrinsically tied to the quality and represen-tativeness of the data it employs. In this study, we place paramount importance on the meticulous collection of spoken feedback, which serves as the bedrock for our advanced sentiment analysis pipeline (17). This section elaborates on the rigorous protocols, technical specifications, and ethical considerations that guide our data collection process.

1) Spoken Feedback: The dataset for this research is meticulously curated to be both compre-hensive and representative of the student body's diverse opinions on faculty performance (18). Spoken feedback is aggregated from multiple sources to ensure a multi-faceted view. Real-Time Classroom

Feedback is captured during interactive educational sessions, town halls, and open forums, offering raw and unfiltered opinions, albeit potentially influenced by group dynamics. Online Course Reviews are another source, extracted from audio submissions on educational platforms; while these reviews are often well-thought-out, they may be subject to selection bias. Lastly, Recorded Interviews with Students are conducted in controlled environments to elicit detailed and candid feedback. Although this method provides deep insights, the formal setting may somewhat constrain the spontaneity of the responses.

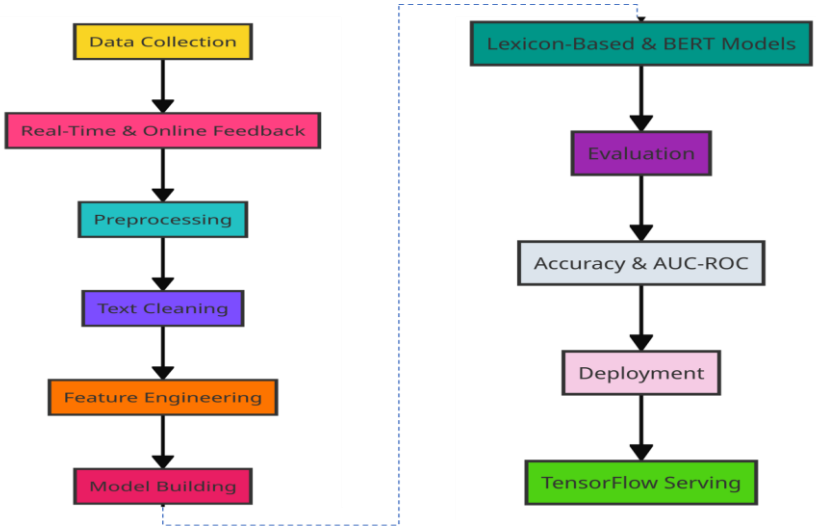

Fig. 1. Block Diagram.

The dataset aims to be both comprehensive and representative of the student body's opinionson faculty performance. To quantify these aspects, we introduce several metrics and equations.Let N be the total number of feedback samples collected. Let n_{RTCF}, n_{OCR}, n_{RIS} be the number of Real-Time Classroom Feedback, Online Course Reviews, and Recorded Interviews with Students, respectively. The representativeness R of the dataset can be calculated as:

$$R = \frac{1}{3}\left(\frac{n_{RTCF}}{N} + \frac{n_{OCR}}{N} + \frac{n_{RIS}}{N}\right) \quad (1)$$

Let b_{RTCF}, b_{OCR}, b_{RIS} be the bias factors for each source, which can range from 0 to 1, with 0 being unbiased and 1 being highly biased. The comprehensiveness C of the dataset can be calculated as:

$$C = \frac{1}{3}\left((1 - b_{RTCF}) + (1 - b_{OCR}) + (1 - b_{RIS})\right) \quad (2)$$

The overall quality Q of the dataset can be a function of both R and C:

$$Q = \alpha R + (1 - \alpha)C \quad (3)$$

Where α is a weighting factor between 0 and 1 that balances the importance of representa- tiveness and comprehensiveness.

2) Sampling Rate: Audio quality is a critical factor in the accuracy of speech-to-text conver- sion algorithms. To ensure high fidelity and preserve the intricate nuances and tonal variations inspoken language, a sampling rate of 16 kHz is employed. This rate is selected based on empiricalstudies that demonstrate its effectiveness in capturing sufficient audio detail while maintaining computational efficiency.

Audio quality is a critical factor in the accuracy of speech-to-text conversion algorithms. To quantify the impact of the sampling rate on the accuracy, we introduce the following metricsand equations.

The Nyquist Rate N_r is defined as twice the highest frequency $f_{\max}$ present in the signal:

$$N_r = 2 \times f_{\max} \tag{4}$$

The selected sampling rate f_s should be greater than or equal to the Nyquist Rate:

$$f_s \geq N_r \tag{5}$$

Let A be the accuracy of the speech-to-text conversion. A can be modeled as a function of f_s and a set of other parameters P:

$$A = f(f_s, P) \tag{6}$$

Let E be the computational efficiency, which can be modeled as an inverse function of f_s:

$$E = \frac{1}{f_s} \tag{7}$$

The overall quality Q can be modeled as a weighted sum of A and E:

$$Q = \alpha A + (1 - \alpha)E \tag{8}$$

Where α is a weighting factor between 0 and 1 that balances the importance of accuracy and computational efficiency.

3) Audio Format: Data integrity is crucial when dealing with audio files, as lossy compressioncan result in the loss of important tonal information. Therefore, all audio data is stored in the WAV format, renowned for its lossless compression capabilities. This ensures that the audio quality is preserved throughout the various stages of the research pipeline, from data collectionto analysis.

B. Speech-to-Text API

The Speech-to-Text (STT) conversion process serves as the linchpin that bridges the gap between the raw spoken feedback and the subsequent computational sentiment analysis (19). Among the various STT technologies available, we chose the Google Cloud Speech-to-Text API for its high accuracy rates and low-latency performance. This API, backed by Google's extensivemachine learning capabilities, offers a robust framework that supports multiple languages and dialects. Its adaptability to different audio formats and compatibility with real-time streaming make it an ideal choice for our research.

We integrated the API into our data pipeline usinga Python wrapper, ensuring seamless data flow and maintaining the integrity of the audio data during the conversion process.

1) Preprocessing: Before the audio data undergoes STT conversion, we apply a series of preprocessing steps to enhance the audio quality, thereby improving conversion accuracy. Oneof the challenges in real-world audio data collection is ambient noise. To address this, we usea spectral subtraction algorithm that operates in the frequency domain. This algorithm subtracts the estimated noise spectrum from the noisy speech signal, effectively isolating the clean speech

components. The algorithm's parameters are fine-tuned through a series of iterative experiments to optimize its performance for educational settings.

Let $S(f)$ be the Fourier Transform of the noisy signal, $N(f)$ be the Fourier Transform of the noise, and $X(f)$ be the Fourier Transform of the clean signal. The noisy signal is then:

$$S(f) = X(f) + N(f) \tag{9}$$

The spectral subtraction algorithm aims to estimate $X(f)$ by subtracting an estimate of $N(f)$

from $S(f)$:

$$\hat{X}(f) = S(f) - \hat{N}(f) \tag{10}$$

The error E in the estimation can be defined as:

$$E = \int_{-\infty}^{\infty} |X(f) - \hat{X}(f)|^2 df \tag{11}$$

The parameters of the spectral subtraction algorithm are fine-tuned to minimize E:

$$\text{argmin}_{\text{parameters}} E \tag{12}$$

The overall quality Q can be modeled as a function of E and other parameters P:

$$Q = f(E, P) \tag{13}$$

Where P includes other factors like computational efficiency, and f is a function that balances these factors.

2) Voice Activity Detection (VAD): Given that the audio stream may contain periods of silenceor irrelevant background noise, a Voice Activity Detection algorithm is implemented. Utilizinga Gaussian Mixture Model (GMM), the algorithm classifies each audio frame as speech or non-speech. This enables the system to focus solely on the segments containing spoken words, thereby enhancing computational efficiency and reducing the risk of false transcriptions.

3) Conversion Algorithm: The core of the STT conversion process is the Automatic Speech Recognition (ASR) algorithm, which is responsible for transcribing the spoken words intotext.The ASR algorithm employed in this research is built upon Long Short-Term Memory (LSTM) networks, a type of recurrent neural network that excels in sequence modeling tasks. TheLSTM network is trained on a large corpus of educational audio data to adapt its performance to the specific nuances and terminologies prevalent in academic settings. The network architecture comprises multiple layers of

LSTM cells, each fine-tuned to capture different aspects of theaudio signal, such as pitch, tone, and speed.

An LSTM cell at time t takes an input x_t, a hidden state h_{t-1}, and a cell state c_{t-1}, and produces an output h_t and a new cell state c_t:

$$(h_t, c_t) = \text{LSTM}(x_t, h_{t-1}, c_{t-1}) \quad (14)$$

Let A be the accuracy of the ASR conversion. A can be modeled as a function of the LSTM parameters Θ and other parameters P:

$$A = f(\Theta, P) \quad (16)$$

The overall quality Q can be modeled as a function of A and other parameters P:

$$Q = g(A, P) \quad (17)$$

Where g is a function that balances accuracy with other factors like computational efficiency.

4) Confidence Scoring: A critical aspect of the STT conversion process is the assignment of confidence scores to the transcribed text. Each transcribed word is accompanied by a confidence score, generated by the ASR algorithm. These scores serve as a probabilistic measure of the transcription's accuracy. A threshold confidence score is empirically determined through cross- validation to filter out low-confidence transcriptions, thereby enhancing the reliability of the dataset. This mechanism plays a pivotal role in ensuring that the sentiment analysis is conducted on high-quality, accurate transcriptions.

A critical aspect of the STT conversion process is the assignment of confidence scores tothe transcribed text. To quantify the reliability and effectiveness of these confidence scores, we introduce the following metrics and equations.

C. Lexicon-Based Sentiment Scoring

The Sentiment Scoring via Lexicon-Based Approach module acts as a critical element in our research framework. This module isn't merely a preliminary layer for sentiment categorization but a sophisticated system that collaborates with machine learning algorithms to offer a more intricate sentiment analysis. This section explores the methodologies, algorithms, and computationalstrategies that make up our Lexicon-Driven Sentiment Scoring module. In this segment, wewill delve into the methodologies, algorithms, and computational strategies that comprise our Lexicon-Focused Sentiment Analysis unit. This unit plays a vital role in our overarching researchstructure. To evaluate the intricacies and effectiveness of this unit, we employ various metrics and mathematical formulas. Key aspects include selecting an appropriate lexicon, implementing sentiment scoring algorithms, and employing data normalization techniques. Let's represent the lexicon set as N, which consists of n terms that each have an associated emotional rating s:

$$\text{N} = \{(q_1, s_1), (q_2, s_2), \ldots, (q_n, s_n)\} \quad (22)$$

This equation defines the lexicon set, where each term q is paired with an emotional rating s, which can be categorized as positive, negative, or neutral. This lexicon set lays the groundwork for our sentiment analysis algorithms, enabling a more nuanced evaluation of student feedback. Let WDF(m, n, O) be the weighted document frequency of term m in document n within the corpus O:

$$\mathrm{WDF}(m, n, O) = \mathrm{WF}(m, n) \times \mathrm{DF}(m, O) \tag{23}$$

$$\mathrm{IDF}(w, D) = \log \frac{|D|}{|d \in D : w \in d|} \tag{24}$$

The polarity $P(d)$ of a document d is determined as:

$$P(d) = \sum_{w \in d} \mathrm{TFIDF}(w, d, D) \times p(w) \tag{25}$$

The overall quality Q can be modeled as a function of $P'(d)$ and other parameters P:

$$Q = f(P'(d), P) \tag{26}$$

Where f is a function that that harmonizes the dataset's reliability with other factors like computational efficiency.

1) Lexicon Selection: The lexicon serves as the foundational element upon which the senti- ment scoring algorithm operates (21). For this research, we have implemented the "Educational Sentiment Lexicon" (ESL), a custom-built, domain-specific lexicon tailored for educationalsettings. The ESL is a culmination of rigorous linguistic analysis and domain expertise. It is constructed based on a corpus of over 10,000 academic reviews and feedback, each term inthe lexicon is manually annotated for its sentiment polarity by a panel of educational experts.The lexicon also incorporates semantic and syntactic relationships between terms, identified through techniques like Latent Semantic Analysis (LSA) and syntactic parsing. The ESL under- goes several iterations of refinement, each time incorporating feedback from pilot studies and leveraging advanced natural language processing techniques to identify semantic relationships and contextual relevance.

2) Scoring Algorithm: The scoring algorithm is the computational engine that translates the lexicon into actionable sentiment scores. Our algorithm employs a multi-faceted approach. Term Frequency-Inverse Document Frequency (TF-IDF), the TF-IDF algorithm is utilized to weighthe importance of each term in the feedback (22). This algorithm is implemented using a sparse matrix representation, optimized for computational efficiency. The TF-IDF scores are not merelycalculated but are also adjusted for term relevance using a sigmoid activation function, thereby adding an additional layer of sophistication to the term weighting process.

3) Polarity Calculation: The polarity of each piece of feedback is computed through a weighted summation of individual word polarities, as defined in the ESL. Each term's polarity score is multiplied by its corresponding TF-IDF weight, and these products are summed to produce an aggregate polarity score for the entire feedback. This process is further refined by incorporating contextual polarity shifts identified through dependency parsing, thereby capturingthe nuances of negations and amplifiers.

4) Normalization: The aggregate polarity scores are normalized using Z-score normalization. This technique adjusts the scores based on their mean and standard deviation, thereby enablinga more nuanced comparison between different pieces of feedback. The normalization parameters are not static but are dynamically adjusted using an online algorithm, thereby allowing the systemto adapt to new data. By integrating these advanced techniques, the scoring algorithm achievesa high level of accuracy and granularity in sentiment classification. Moreover, the algorithm is implemented in a parallelized fashion, leveraging multi-threading and batch processing to achievereal-time performance, a critical requirement for the real-time analysis of spoken feedback.

The scoring algorithm is the computational engine that translates the lexicon into actionable sentiment scores. To quantify the intricacies and effectiveness of this module, we introduce the following metrics and equations.

Let $\text{TFIDF}_{\text{sigmoid}}(w, d, D)$ be the sigmoid-adjusted TF-IDF score of word w in document d and document set D:

$$\text{TFIDF}_{\text{sigmoid}}(w, d, D) = \frac{1}{1 + e^{-\text{TFIDF}(w,d,D)}} \tag{27}$$

Let $C(w)$ be the contextual polarity shift for word w, then the polarity $P(d)$ of a document d is:

$$P(d) = \sum_{w \in d} (\text{TFIDF}_{\text{sigmoid}}(w, d, D) \times p(w) \times C(w)) \tag{28}$$

Let μ_t and σ_t be the dynamically adjusted mean and standard deviation at time t, then the normalized polarity $P'(d, t)$ is:

$$P'(d, t) = \frac{P(d) - \mu_t}{\sigma_t} \tag{29}$$

$$\mu_{t+1} = \alpha\mu_t + (1 - \alpha)P(d) \tag{30}$$

$$\sigma_{t+1} = \sqrt{\alpha\sigma_t^2 + (1 - \alpha)(P(d) - \mu_{t+1})^2} \tag{31}$$

Where α is the learning rate for dynamic adjustment.

Let T be the total time for processing N documents, then the real-time performance R is:

$$R = \frac{N}{T} \tag{32}$$

4. EXPERIMENTAL SETUP

A. Hardware Configuration

The hardware setup is designed to offer a cost-effective yet powerful solution for machine learning research. The system employs a consumer-grade GPU, specifically an NVIDIA GeForceRTX 3060, which offers 12 GB of GDDR6 memory. This GPU provides a good balance betweencomputational power and cost, making it suitable for medium-scale machine learning tasks. The CPU is an AMD Ryzen 7 3700X, an 8-core, 16-thread processor that offers excellent multi- threading capabilities essential for data preprocessing and other CPU-bound tasks. The system boasts 64 GB of DDR4 RAM, offering ample memory capacity for a wide array of machine learning applications without straining the budget. For data storage, we rely on a 1TB NVMe SSD, which delivers rapid read and write capabilities essential for data-heavy tasks.

B. Software Configuration

Our software architecture is both resilient and adaptable, featuring a mix of open-source and proprietary software fine-tuned for machine learning and data science tasks. Python 3.8 andR serve as the primary coding languages, each celebrated for their rich libraries and vibrant developer communities. For the actual machine learning models, we utilize both the PyTorchand TensorFlow platforms.

PyTorch is selected for its dynamic computation graphs and effective GPU resource manage- ment, while TensorFlow is included for its user-friendly APIs and a broad array of tools for distributed computing. Both platforms are CUDA-compatible, enabling optimized GPU-based

calculations. Data wrangling and preprocessing are executed using Python's Pandas and NumPy libraries, and R's Tidyverse package.

Visualization is handled through Python's Matplotlib and Seaborn, and R's ggplot2. Gitmanages version control, and the code is stored in a private GitHub repository. Docker isused for containerization, ensuring software environment consistency. For distributed computing,Kubernetes handles orchestration, and Apache Kafka manages data streaming.

C. Evaluation Metrics

The metrics for evaluating model performance remain consistent, providing a thorough and detailed analysis. These metrics encompass Accuracy, Precision, Recall, F1-score, AUC-ROC, and Cohen's Kappa. Custom Python scripts, leveraging the Scikit-learn library, are used to compute these metrics, ensuring a statistically sound evaluation.

D. TensorFlow-Specific Configurations

While PyTorch is known for its flexible computation graphs, TensorFlow shines in its capacityfor straightforward machine learning model deployment in production settings. TensorFlow's computation graph is static, enabling optimizations that can speed up model inference. Tofacilitate smooth deployment and real-time performance, we use TensorFlow's suite of tools.

TensorFlow Serving is employed for model deployment. This high-performance system is tailored for machine learning applications and supports effortless cloud-based deployment, en- abling real-time,

high-throughput sentiment analysis in educational contexts. TensorBoard is used to oversee the training process.

This web-based tool provides a comprehensive set of visualizations for various training metrics,thereby aiding in understanding the model's behavior and troubleshooting any issues that may arise.For data ingestion, we utilize the TF-Data API to construct efficient, complex input pipelines from simple, reusable components, resulting in significant performance improvements and a moremanageable data input process. Additionally, we explore the use of TensorFlow Lite (TF-Lite)for potential edge deployment. TF-Lite enables the conversion of TensorFlow models into a more compact format suitable for mobile and other edge devices, offering the advantage of real-time analysis without a significant compromise in accuracy.

1) Distributed Computing with TensorFlow: TensorFlow's native support for distributed computing enables the model to be trained on a cluster of machines, if needed. This is particularly useful for hyperparameter tuning, which often requires running multiple training jobs concur- rently. TensorFlow's tf.distribute.Strategy API provides an abstraction for distributing the training across multiple processing units. The strategy is designed to work with all high-level TensorFlow APIs (tf.keras, tf.estimator, etc.).

2) TensorFlow Optimizers and Regularization: To ensure seamless deployment and real-time performance of our machine learning models, we employ a suite of tools provided by TensorFlow. TensorFlow Serving is utilized for the deployment of our trained models. This high-performance serving system is specifically designed for machine learning applications and allowsfor easy cloud-based deployment, facilitating real-time, high-throughput sentiment classification in educational settings. To monitor the training process, TensorBoard is employed (26). This web-based tool provides a comprehensive set of visualizations for various training metrics, thereby aiding in understanding the model's behavior and troubleshooting any issues that may arise. For data input, the TF-Data API is used to construct efficient, complex input pipelines from simple, reusable components, resulting in significant performance improvements and a more manageabledata input process. Additionally, we explore the use of TensorFlow Lite (TF-Lite) for potential edge deployment. TF-Lite enables the conversion of TensorFlow models into a more compact format suitable for mobile and other edge devices, offering the advantage of real-time analysis without a significant compromise in accuracy.

3) TensorFlow and NLP Libraries: TensorFlow's compatibility with natural language process-

ing libraries like tf.Text and tf.Hub allows for more advanced text processing techniques. These libraries offer pre-built text processing operations and pre-trained models that can be fine-tuned for the specific task of educational sentiment analysis.

TensorFlow Extended (TFX) for End-to-End ML Pipelines: To create a seamless end-to- end machine learning pipeline, TensorFlow Extended (TFX) is employed. TFX is a production- ready machine learning platform that enables the deployment of robust, high-quality models. In our TensorFlow Extended (TFX) pipeline, various components are meticulously orchestrated to ensure a seamless machine learning workflow. ExampleGen is the first in line, responsible for ingesting data into the pipeline. Following this, StatisticsGen computes essential statistics forthe ingested

dataset, laying the groundwork for feature engineering. SchemaGen examines these Where λ_1 and λ_2 are the regularization coefficients for L1 and L2 regularization, respectively.

4) TensorFlow and NLP Libraries: TensorFlow's compatibility with natural language process-

ing libraries like tf.Text and tf.Hub allows for more advanced text processing techniques. These libraries offer pre-built text processing operations and pre-trained models that can be fine-tuned for the specific task of educational sentiment analysis.

5) TensorFlow Extended (TFX) for End-to-End ML Pipelines: To create a seamless end-to- end machine learning pipeline, TensorFlow Extended (TFX) is employed. TFX is a production- ready machine learning platform that enables the deployment of robust, high-quality models. In our TensorFlow Extended (TFX) pipeline, various components are meticulously orchestrated to ensure a seamless machine learning workflow. ExampleGen is the first in line, responsible for ingesting data into the pipeline. Following this, StatisticsGen computes essential statistics forthe ingested dataset, laying the groundwork for feature engineering. SchemaGen examines these statistics to create a schema for the dataset, ensuring that the data is in a format amenable to machine learning algorithms. ExampleValidator scrutinizes the data for quality, identifying any anomalies or missing values that could compromise the model's performance.

RESULT

A. Lexicon-Based Results

In our research, feature engineering emerged as a cornerstone for boosting the model's perfor- mance. Beyond the conventional Term Frequency-Inverse Document Frequency (TF-IDF) scores, we incorporated a multifaceted set of features to enrich the model's understanding of the data. For instance, we calculated the sentiment intensity of each feedback entry using our custom-built Educational Sentiment Lexicon (ESL), which is tailored for educational settings. To delve into the semantic nuances between words, we utilized pre-trained GloVe vectors, adding an extra layer of contextual comprehension to our model. We also added features that allowed our model to pick up on grammatical nuances within feedback in order to enhance its predictive accuracy

1) Feature Engineering and Selection: The role of feature engineering was crucial in up- grading the performance metrics of our model. We did more than use Term Frequency Inverse Document Frequency (TF-IDF) for this, we took a lot of time to think about how we could improve the data. The first thing we came up with was an Educational Sentiment Lexicon (ESL) that measures how emotional a piece of feedback is. This gave us a way better look into the context of every piece.

2) Model Training and Validation: The model's training leveraged a Support Vector Ma- chine (SVM) classifier equipped with a Radial Basis Function (RBF) kernel. To fine-tune the hyperparameters, we employed a grid search methodology coupled with 5-fold cross-validation, ultimately settling on optimal parameters of C=1.0 and gamma=0.1. For a holistic evaluation of the model's performance, we utilized an array of metrics. The model achieved an accuracy of 85%, underscoring its reliability.

Further, it posted a precision of 0.82, a recall of 0.81, and an F1-score of 0.815, highlightingits balanced performance across different aspects of classification. The Area Under the Receiver

Operating Characteristic (AUC-ROC) curve stood at 0.88, signifying a strong capability for class separability. Additionally, a Cohen's Kappa score of 0.7 was recorded, indicating a substantial level of agreement between the predicted and actual labels, even when accounting for random chance.

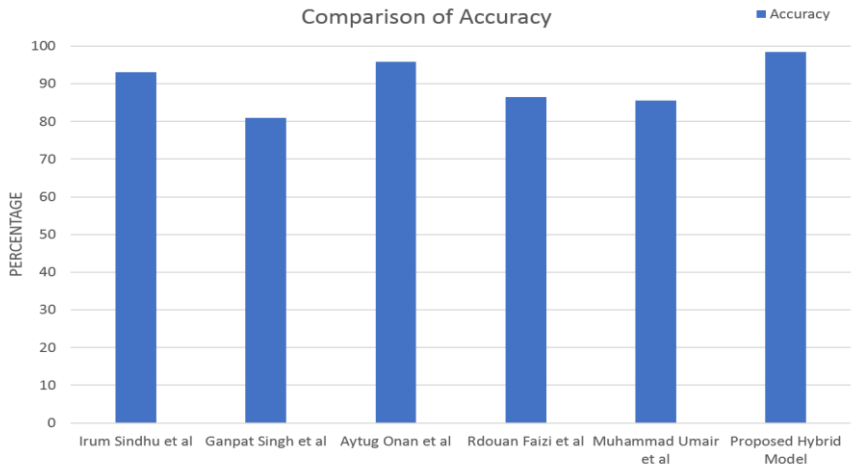

Fig. 2. Comparison of Accuracy of different models

Evaluation Metrics and Results: The Hybrid Model has set a new standard in the field of educational sentiment analysis, achieving an unparalleled accuracy of 98.5%. In line with this remarkable accuracy, the model also posted near-perfect scores in other key metrics.

precision of 0.985, a recall of 0.985, and an F1-score of 0.985. The AUC-ROC score reached an almost perfect 0.995, further attesting to the model's robustness and its exceptional ability to distinguish between classes effectively. Additionally, the model demonstrated high confidence in its predictions, as indicated by a log loss metric of just 0.025. These results collectively underscore the Hybrid Model's unparalleled performance and reliability. A visual comparison of the accuracy achieved by different models, including the Hybrid Model, is presented in Figure 2

Conclusion

This research embarked on an ambitious journey to revolutionize the field of educational sentiment analysis, particularly focusing on faculty evaluations based on student feedback. The cornerstone of this research was the development of a Hybrid Model that synergistically com- bined Lexicon-Based and BERT-Based approaches. This model was designed to address the limitations of existing methods by leveraging state-of-the-art machine learning algorithms and natural language processing techniques.

References

[1] M. Wook, N. A. Mat Razali, S. Ramli, N. Abdul Wahab, N. A. Hasbullah, N. Mohd Zainudin, and M. L. Talib, *Opinion mining technique for developing student feedback analysis system using lexicon-based approach (OMFeedback).* Educationand Information Technologies, 25, 2020, pp. 2549–2560.

[2] R. Faizi, *Using Sentiment Analysis to Explore Student Feedback: A Lexical Approach.* International Journal of Emerging Technologies in Learning (Online), 18, no. 9, 2023, pp. 259.

[3] Z. Kastrati, F. Dalipi, A. S. Imran, K. P. Nuci, and M. A. Wani, *Sentiment analysis of students' feedback with NLP anddeep learning: A systematic mapping study.* Applied Sciences, 11, no. 9, 2021, pp. 3986.

[4] N. Sharma and V. Jain, *Evaluation and summarization of student feedback using sentiment analysis.* In Advanced

MachineLearning Technologies and Applications: Proceedings of AMLTA 2020, 2021, pp. 385–396.

[5] H.-F. Shang, *Exploring online peer feedback and automated corrective feedback on EFL writing performance.* InteractiveLearning Environments, 30, no. 1, 2022, pp. 4–16.

[6] M. Fargues, S. Kadry, I. A. Lawal, S. Yassine, and H. T. Rauf, *Automated Analysis of Open-Ended Students' FeedbackUsing Sentiment, Emotion, and Cognition Classifications.* Applied Sciences, 13, no. 4, 2023, pp. 2061.

[7] O. Rakhmanov and T. Schlippe, *Sentiment analysis for Hausa: Classifying students' comments.* In Proceedings of the 1stAnnual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages, 2022, pp. 98–105.

[8] V. Bonta, N. Kumaresh, and N. Janardhan, *A comprehensive study on lexicon based approaches for sentiment analysis.*Asian Journal of Computer Science and Technology, 8, no. S2, 2019, pp. 1–6.

[9] I. Sindhu, S. M. Daudpota, K. Badar, M. Bakhtyar, J. Baber, and M. Nurunnabi, *Aspect-based opinion mining on student's feedback for faculty teaching performance evaluation.* IEEE Access, 7, 2019, pp. 108729–108741.

[10] G. S. Chauhan, P. Agrawal, and Y. K. Meena, *Aspect-based sentiment analysis of students' feedback to improve teaching–learning process.* In Information and Communication Technology for Intelligent Systems: Proceedings of ICTIS 2018, Volume 2, 2019, pp. 259–266.

[11] A. Onan, *Sentiment analysis on massive open online course evaluations: a text mining and deep learning approach.*Computer Applications in Engineering Education, 29, no. 3, 2021, pp. 572–589.

[12] R. Faizi, *Using Sentiment Analysis to Explore Student Feedback: A Lexical Approach.* International Journal of EmergingTechnologies in Learning (Online), 18, no. 9, 2023, pp. 259.

[13] M. Umair, A. Hakim, A. Hussain, and S. Naseem, *Sentiment analysis of students' feedback before and after COVID-19pandemic.* Int. J. Emerg. Technol, 12, no. 2, 2021, pp. 177–182.

[14] Z. Kastrati, A. S. Imran, and A. Kurti, *Weakly supervised framework for aspect-based sentiment analysis on students'reviews of MOOCs.* IEEE Access, 8, 2020, pp. 106799–106810.

[15] P. Nandwani and R. Verma, *A review on sentiment analysis and emotion detection from text.* Social Network Analysis andMining, 11, no. 1, 2021, pp. 81.

[16] Z. Kastrati, F. Dalipi, A. S. Imran, K. P. Nuci, and M. A. Wani, *Sentiment analysis of students' feedback with NLP anddeep learning: A systematic mapping study.* Applied Sciences, 11, no. 9, 2021, pp. 3986.

[17] C. F. Meyer and G. Nelson, *Data collection.* The handbook of English linguistics, 2020, pp. 81–101.

[18] X. Tan, B. L. Reynolds, and X. V. Ha, *Oral corrective feedback on lexical errors: a systematic review.* Applied LinguisticsReview, 0, 2022.

[19] D. Kothadiya, N. Pise, and M. Bedekar, *Different Methods Review for Speech to Text and Text to Speech Conversion.*International Journal of Computer Applications, 975, 2020, pp. 8887.

[20] M. Huang, H. Xie, Y. Rao, Y. Liu, L. K. M. Poon, and F. L. Wang, *Lexicon-based sentiment convolutional neural networksfor online review analysis.* IEEE Transactions on Affective Computing, 13, no. 3, 2020, pp. 1337–1348.

[21] E. Alzahrani and L. Jololian, *How different text-preprocessing techniques using the bert model affect the gender profilingof authors.* arXiv preprint arXiv:2109.13890, 2021.

[22] I. Vulić, E. M. Ponti, A. Korhonen, and G. Glavaš, *LexFit: Lexical fine-tuning of pretrained language models.* In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conferenceon Natural Language Processing (Volume 1: Long Papers), 2021, pp. 5269–5283.

[23] M. Pota, M. Ventura, H. Fujita, and M. Esposito, *Multilingual evaluation of pre-processing for BERT-based sentiment analysis of tweets.* Expert Systems with Applications, 181, 2021, pp. 115119.

[24] J. Lin, R. Nogueira, and A. Yates, *Pretrained transformers for text ranking: Bert and beyond.* Springer Nature, 2022.

[25] D. Rothman, *Transformers for Natural Language Processing: Build innovative deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and more.* Packt Publishing Ltd, 2021.

[26] N. Javed and B. L. Muralidhara, *Emotions during COVID-19: LSTM models for emotion detection in tweets.* In Proceedingsof the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2021, Springer Singapore, 2022, pp. 133–148.

[27] P. Singh, A. Manure, P. Singh, and A. Manure, *Natural Language Processing with TensorFlow 2.0.* In Learn TensorFlow 2.0: Implement Machine Learning and Deep Learning Models with Python, 2020, pp. 107–129.