The New England Journal of Statistics in Data Science Volume 0, 1–18 (2026)
DOI: https://doi.org/10.51387/26-NEJSDS105

Comparative Analysis of NLP Methods for Emotion Detection in Student Responses During COVID-19

Alexander MARET, Cade DEES, Yule FU, Yanjun QIAN*, David CHAN, Punit GANDHI, and Indranil SAHOO

Abstract

Natural language processing (NLP) algorithms have demonstrated significant capabilities in understanding responses to open-ended questions in survey data. However, the reliability and uncertainty of these methods on this task still need to be thoroughly investigated. To address this issue, this paper presents a comprehensive comparative analysis of various NLP methods for detecting fine-grained emotions in student responses about their mental health during the COVID-19 pandemic. The evaluated models include a Lexicon-based approach, the bag-of-words (BoW) model, Term Frequency-Inverse Document Frequency (TF-IDF), a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model, MentalBERT, and OpenAI's GPT-3.5. We carefully assess the efficacy of these models in accurately classifying emotions into predetermined categories using performance metrics such as accuracy and F1 score. Furthermore, model stability and distinguishing ability were quantified through repetitive cross-validation and the Area Under the Receiver Operating Characteristic Curve (AUC). The consistency of emotion detection across different models is also evaluated. The study highlights that the effectiveness of employing NLP methods for mental health analysis may vary depending on the emotions being analyzed, and their stability and uncertainty require thorough examination. Our work can provide valuable guidance for data scientists on applying NLP methods to survey data, particularly for understanding survey respondents' emotions.

KEYWORDS AND PHRASES: Survey responses, COVID-19, Natural language processing, Emotion detection.

1. INTRODUCTION

In the early 2020s, the lives of university students were significantly impacted by the COVID-19 pandemic. The abrupt transition to online learning, loss of contact with family and friends, and social and financial uncertainty had a considerable impact on students' mental health. Many surveys [35, 42, 9, 18] have been conducted in colleges and universities to assess their impact and long-term effects. These surveys typically included two types of questions: quantitative questions that measure agreement or disagreement with specific statements, and open-ended questions that invite students to describe their personal feelings and experiences. Traditional survey analysis [24] often focuses on the quantitative responses, as these numerical responses can be easily analyzed using mathematical and statistical models. However, the responses to open-ended questions can provide more details and personalized insights into students' mental health, which can then lead to studying the reasons behind these impacts. To analyze this information in large data sets, it is necessary to apply Natural Language Processing (NLP) techniques [23] to detect emotions within the students' responses.

Advancements in NLP for emotion detection have been significant over the past few decades. Early researchers developed the Lexicon-based method [38] to create a dictionary that scores the emotions associated with individual words. Subsequently, techniques like the Bag-of-Words (BoW) model [29] and Term Frequency-Inverse Document Frequency (TF-IDF) [32] were introduced to convert texts into numerical vectors. Machine learning algorithms, such as logistic regression [36] and Support Vector Machines (SVM) [34], can then be trained using these vectors to classify texts into different emotion categories. Since 2017, the Transformer architecture [41] in deep learning has achieved remarkable success in NLP. Large Language Models (LLMs) [26], such as the Bidirectional Encoder Representations from Transformers (BERT) [12] and the Generative Pre-trained Transformer (GPT) [30], have demonstrated strong capabilities in various NLP tasks. We can either use these LLMs directly for emotion detection or fine-tune them for specific emotions to further enhance accuracy.

Despite the success of NLP methods, two significant challenges remain when applying them to understand the mental health of college students through open-ended survey questions. First, mental health encompasses a range of nuanced emotions, such as depression, anxiety, stress, and iso-

*Corresponding author.

lation. However, many traditional sentiment analysis methods [24] only consider general positive or negative sentiments in texts. This limitation can lead to the neglect of personal emotions expressed in open-ended survey responses, making it difficult to identify subtle differences in mental health. Second, using NLP to analyze mental health can yield educational and psychological insights; thus, the stability and distinguishing ability of these methods are crucial for reliable studies. There has been some effort to explore differences between such methods in the context of social media text [4, 25] and free response surveys [44, 43, 19] related to COVID. However, most of those studies only focus on the consistency in the general sentiment prediction, not on fine-grained emotion detection [10] that we are concerned with.

To address the challenges of reliable emotion detection, we conduct a comparative analysis of various NLP methods based on a recent study of college student responses during COVID-19 [1]. This study surveyed students at a large mid-Atlantic university in the U.S. during the early months of the COVID-19 pandemic to assess its impact on their mental health. The survey included both traditional quantitative scoring questions and open-ended responses. In their research, Amona et al. [1] carefully annotated ten common emotions—such as isolation, depression, and anxiety—derived from the students' responses to the question, "How is COVID affecting your mental health?" They then examined how these emotions impacted different subgroups within the student population.

Here, we compare a wide range of NLP methods, including Lexicon-based approaches, BoW, TF-IDF, fine-tuned BERT, and zero-shot GPT, for the automatic detection of emotions in these survey responses. First, we assess the performance of various NLP methods across all identified emotions. We find that despite the complicated Transformer method achieving the best overall performance, simpler methods, such as Lexicon, can effectively identify specific emotions. Next, we evaluate the stability and distinguishing ability of these models, demonstrating the performance-complexity trade-off when applying NLP methods. Finally, we evaluate the detection consistency in emotion detection between NLP methods and the true labels, assessing whether these methods yield similar results or not. This comprehensive study highlights that the effectiveness of employing NLP methods to analyze mental health through survey data varies for the emotions being analyzed. Moreover, for the methods with top overall performance, their stability and uncertainty need to be thoroughly examined. We summarize insights from our experimental studies and offer method selection recommendations for NLP analysis of survey data to guide future data science practices.

The paper is organized as follows: In Section 2, we introduce related work that examines students' mental health during COVID-19 and advancements in NLP methods for emotion detection. Section 3 outlines three aspects of our methodology: data collection and annotation, the implementation of NLP methods, and the comparison framework. In Section 4, we present our results and discuss the outcomes of emotion detection using NLP methods in relation to mental health. Finally, Section 5 summarizes the conclusions of our study.

2. RELATED WORK

2.1 NLP Methods in Comparison Study

First, we will review the related NLP methods in our comparison study. Traditional methods include the Lexicon-based approach, where predefined dictionaries of emotional words are used to identify emotions in text. For example, Mohammad et al. [21] developed the National Research Council of Canada (NRC) Emotion Lexicon (EmoLex), a widely used resource for Lexicon-based emotion detection. The model establishes connections between words and basic emotions, including anger, joy, and sadness. This Lexicon was developed through crowdsourcing, ensuring a diverse and comprehensive set of word-emotion associations. Mohammad [20] later extended their work by adding real-valued scores of intensity to emotions to create NRC Affect Intensity Lexicon (AIL), enabling more fine-grained analysis.

The next school of methods for text classification involves converting sentences to numeric vectors using BoW or TF-IDF and applying machine learning algorithms to them. Sebastiani [33] provided a thorough analysis of these algorithms, highlighting their performance across different datasets and establishing their strengths and limitations in text classification. BoW is a technique that turns text or images into a histogram of words. BoW models convert text into a matrix of token counts, representing the frequency of each word in the text. This representation is then used as input for machine learning classifiers such as logistic regression, SVM, or Naive Bayes. Based on the study in [29], this makes the BoW computationally simple, helping it score well on performance tests. Barry [2] studied using BoW on Amazon and Yelp food reviews to classify whether they were positive or negative. With its best machine learning model, they achieved an accuracy score of over 95%. Desmet and Hoste [11] used BoW to detect 15 emotions. Their results varied by emotion, but six of the seven most common emotions had acceptable accuracy.

TF-IDF improves upon BoW by weighting terms based on their importance, calculated as the product of Term Frequency (TF) and Inverse Document Frequency (IDF). Ramos et al. [32] explained that the less frequently a word appears in documents, the greater the weight it should receive. This weighting helps to emphasize significant words while downplaying common ones, enhancing the model's ability to distinguish between different classes. Rahman et al. [31] conducted sentiment classification by tweaking TF-IDF with various vectorization methods and classifiers. With

the correct classifier, they achieved 100% accuracy. Sundaram et al. [37] used TF-IDF for six emotions. For emotions with large training sets, they had an accuracy of about 85%.

The advent of the Transformer architecture in deep learning, such as BERT and GPT, has revolutionized NLP. BERT, introduced in [12], employs a bidirectional training approach to understand the context of words in a sentence. BERT's architecture comprises multiple layers of encoders within the Transformer, enabling it to capture intricate relationships between words. BERT can be fine-tuned for specific tasks, such as emotion detection, which involves additional training on a labeled dataset to optimize the model's performance for that particular task. Tang et al. [39] further explored the fine-tuning of BERT for multi-label sentiment analysis, showcasing its effectiveness in handling multiple co-occurring emotions under unbalanced class distributions. Ji et al. [17] developed MentalBERT, a BERT-based model fine-tuned on mental health-related text, demonstrating significant improvements in understanding and classifying emotional content compared to standard BERT models.

GPT models [30, 27], such as GPT-3.5, leverage generative pre-training on a vast corpus of text to generate human-like responses. Floridi and Chiriatti [13] explained that such models will transform the writing process and are capable of producing texts on the level of some humans. These models can be adapted for emotion detection by fine-tuning them on specific datasets or using prompt engineering to elicit desired outputs. Jain et al. [16] used two GPT models for emotion detection, achieving an accuracy score of 0.98 over the mental health datasets they tested it on. The BERT and GPT models show much promise, with the GPT models being the most cutting-edge technology available.

2.2 Fine-Grained Emotion Detection

For this study, multiple fine-grained emotions related to mental health, such as isolation, anxiety, and depression, need to be detected from students' responses to the open-ended question. Bouzazizi et al. [5] tackled the challenging task of multi-class emotion detection on Twitter posts, achieving 60.2% accuracy for seven emotion classes. Their study emphasized the complexity of multi-class classification and proposed a model to better extract and understand emotions present in text rather than classifying them into predefined categories. The authors introduced a system that first classifies text as positive or negative and then assigns scores for corresponding emotion subclasses, improving the robustness and accuracy of emotion classification. Demszky et al. [10] created a labeled dataset of $58k$ comments for 27 emotions, including gratitude, confusion, and remorse. They also trained a BERT-based model, achieving 0.46 F1 score. Mustafa et al. [22] leveraged Twitter data and machine learning to classify depression severity, achieving 91% accuracy by analyzing the top 100 words used by individuals and their psychological attributes. The study highlighted the importance of feature selection in enhancing classifier performance and proposed incorporating additional data, such as emojis and images, to improve future analyses. Guo et al. [14] introduced a multi-way matching deep neural network model for fine-grained emotion detection of user reviews. Their approach predicted scores for specific attributes within reviews, such as location, service, price, and environment. The model consists of two steps: attribute detection and attribute classification. In the first step, the model identifies the relevant attributes mentioned in the text. In the second step, it assigns a score ranging from -5 to 5 for each emotion, reflecting the user's opinion. This fine-grained analysis offers a more detailed understanding of user emotions by focusing on specific aspects of their reviews, demonstrating that NLP methods can effectively distinguish between various emotion categories.

3. METHODOLOGY

3.1 Data Collection and Labeling

The dataset has been previously studied in [1] and [7], and was collected from students at a large mid-Atlantic university in the U.S. between April and June 2020. We focus on one part of the collected data containing short-answer responses from students concerning the impact of COVID-19 on their mental health. The students had to answer the question, "How is coronavirus/COVID-19 affecting your mental health?" The responses are labeled manually by our research team with various emotional indicators, which serve as the ground truth for our emotion detection models. Each response is annotated with binary labels for 10 emotions: *Isolation*, *Depression*, *Anxiety*, *Negative Feelings*, *Lack of Motivation*, *All Stress*, *Issues With Home Life*, *No/Positive Effects*, *Lack of Routine*, *Miscellaneous*. The binary labels (1 if positive, 0 if negative) indicate whether the labelers believed the respondent's answers expressed the corresponding emotions.

The ten categories of emotions are inspired by [6] and [35]. In Appendix A, we provide an example for each emotion, and explain our definitions of *Negative Feelings*, *All Stress*, and *Miscellaneous*. A response could be labeled into more than one emotion category. To ensure the labeling quality, two team members collaboratively categorized emotions for each response. For any questionable answers, they would sort out with multiple members to reach a consensus on labeling. In preparation for our analysis, the data were cleaned by removing responses with no emotion detected, which typically occurred when there was no response or only random characters. We also removed responses that lacked demographic information to facilitate future analysis. This left 398 responses in the dataset. The percentages of the remaining responses labeled as 1 for each emotion, ordered from largest to smallest, are illustrated in Figure 1.

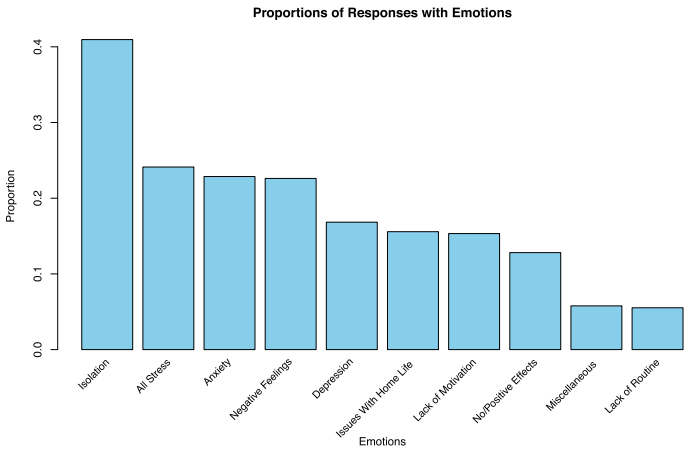

Figure 1: The proportions of the responses expressing the corresponding emotions from the human labeling results. The ten emotions are ordered from the largest to the smallest.

In Figure 2, we analyze correlation and hierarchical clustering for all emotions. We find that most of the emotion pairs have near-zero correlations. The clustering analysis shows that "Depression" and "Anxiety" are the closest emotions. However, their correlation is only 0.3. Other close emotions also show small correlation coefficients of ~ 0.1. Thus, we simplify this multiple-label classification problem into 10 binary classification problems. For every NLP method, we train ten models, each using labels for a single emotion. This simplification will provide a fair method-comparison framework.

3.2 Emotion Detection Using NLP

3.2.1 Text Preprocessing

We follow the common steps in NLP [23] to preprocess the students' responses in all the following methods, except GPT, which takes the original text as input. The preprocessing steps include:

- Text cleaning: Removal of special characters, numbers, and extraneous whitespace.
- Tokenization: Splitting text into individual words or tokens.
- Lowercasing: Converting all text to lowercase to ensure uniformity.
- Stop words removal: Removing common words that do not contribute to emotional meaning, such as "and," "the," etc.
- Lemmatization: Reducing words to their base or root form.

After the preprocessing, the tokenized text will serve as input to the following NLP models to detect emotions expressed in the responses.

3.2.2 Lexicon-Based Method

The Lexicon-based model uses a custom dictionary created from a human-encoded text dataset. After the preprocessing, the word frequencies are calculated to understand the distribution of terms within the dataset. Words are then scored based on their association with emotion labels, using metrics such as pointwise mutual information (PMI) to quantify the strength of association between words and emotions.

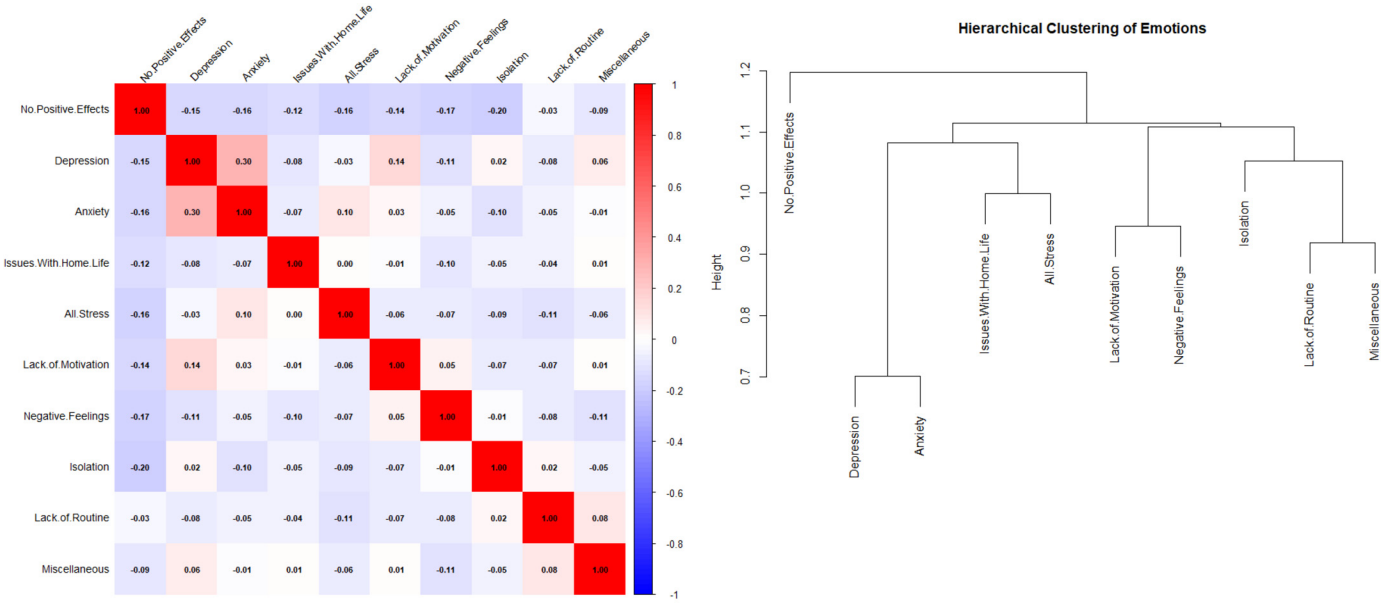

Figure 2: The correlation matrix (left) and dendrogram clustering (right) of the labels for ten emotions.

For a response r with words $w_1, w_2, \ldots, w_n$, its probability of including the emotion $\mathbf{E}_j$ is:

$$p(r, \mathbf{E}_j) = \sigma\{\sum_{i=1}^{n} \text{Score}(w_i, \mathbf{E}_j) + \text{Intercept}(\mathbf{E}_j)\} \quad (3.1)$$

where $\text{Score}(w_i, \mathbf{E}_j)$ represents the score of word w_i for emotion $\mathbf{E}_j$, $\text{Intercept}(\mathbf{E}_j)$ is the intercept for emotion $\mathbf{E}_j$, and $\sigma\{\cdot\}$ is the Sigmoid function. The Lexicon model will conclude that a response expresses an emotion when its predicted probability is at least 0.50.

Our study implemented the Lexicon-based model using R's `SentimentAnalysis` package [28]. All responses were scanned to create a custom dictionary for each emotion. The scores of top words for the three typical emotions identified in Section 4.1, as well as their intercepts, are presented in Table 1. We find that most of the words with coefficients different from 0 are explainable. For the *"Depression"* emotion, "depress" is associated with positive instances, while "routin", "tend", and "schedul" are associated with negative instances. For *"Lack of Motivation"*, "motiv", "focus". and "bed" are associated with positive ones. For *"Miscellaneous"*, the method only finds two words. The word "sleep" shows a clear positive association, indicating the labelers put the sleep issues in this category. On the other hand, the words with near-zero coefficients are less explainable. They might be introduced in the dictionaries due to random sampling of positive/negative instances.

Table 1. The word scores and intercepts of three typical emotions in the customized dictionary for our dataset.

Depression		*Lack of Motivation*		*Miscellaneous*	
Word	**Score**	**Word**	**Score**	**Word**	**Score**
depress	0.321	motiv	0.138	sleep	0.028
routin	−0.033	focus	0.075	focus	0.003
tend	−0.023	bed	0.044		
schedul	−0.016	there	0.020		
becom	−0.005	anymore	0.012		
effect	−0.003	cant	0.001		
Intercept	0.036	Intercept	0.086	Intercept	0.059

3.2.3 BoW and TF-IDF

The BoW [29] and TF-IDF [32] methods will convert the tokenized text into a vector or matrix, and then train machine learning models for emotion detection. The BoW model transforms text into a matrix of token counts. Each response is represented as a vector $\mathbf{v}$ indicating the frequency of each word in the text. For a response r with words $w_1, w_2, \ldots, w_n$, the vector representation $\mathbf{v}_{\text{BoW}}(r)$ is given by:

$$\mathbf{v}_{\text{BoW}}(r) = [f(w_1, r), f(w_2, r), \ldots, f(w_n, r)] \quad (3.2)$$

where $f(w_i, r)$ is the frequency of word w_i in response r.

The TF-IDF model improves upon the BoW model by weighing terms based on their importance. The term frequency (TF) measures how often a word appears in a document, while the inverse document frequency (IDF) measures how unique or rare a word is across all documents. The TF-IDF score for a word w in response r is calculated as:

$$\text{TF-IDF}(w, r) = \text{TF}(w, r) \times \text{IDF}(w) \quad (3.3)$$

where:

$$\text{TF}(w, r) = \frac{f(w, r)}{\sum_{w' \in r} f(w', r)} \quad (3.4)$$

and

$$\text{IDF}(w) = \log\left(\frac{N}{|\{r \in R : w \in r\}|}\right) \quad (3.5)$$

where $f(w, r)$ is the frequency of word w in response r, N is the total number of responses, and $|\{r \in R : w \in r\}|$ is the number of responses containing the word w. Finally, for a response r with words $w_1, w_2, \ldots, w_n$, the vector representation $\mathbf{v}_{\text{TF-IDF}}(r)$ is given by:

$$\mathbf{v}_{\text{TF-IDF}}(r) = [\text{TF-IDF}(w_1, r), \text{TF-IDF}(w_2, r), \ldots, \text{TF-IDF}(w_n, r)].$$

After transforming each response into a vector, we adopt machine learning methods to train classifiers to detect each emotion. In this study, we consider two methods: logistic regression [36] and SVM [34]. Logistic regression predicts the probability of a class by applying a logistic function to a linear combination of input features, whereas SVM finds the hyperplane that best separates the data into classes by maximizing the margin between the classes. With the kernel method [15], those linear classifiers can be extended for non-linear classification. However, the performance of non-linear classifiers depends on the careful choice of kernels and their hyper-parameters for specific problems and datasets. To avoid excessive parameter tuning, we only consider the linear classifiers in the BoW and TF-IDF methods. To handle imbalanced data, we can also use the Synthetic Minority Over-sampling Technique (SMOTE) [8] to generate synthetic samples for the minority class, balancing the dataset. After conducting preliminary experiments, we employ the SVM with a linear kernel for the BoW method and logistic regression with SMOTE for the TF-IDF method, as these combinations provide generally better accuracy across different emotions.

3.2.4 Fine-Tuned MentalBERT

BERT [12] is a transformer-based model designed to understand the context of words in a sentence through bidirectional training. MentalBERT [17] is a pre-trained BERT model specialized in mental health-related text. We first load the pretrained network "mental-bert-base-uncased".

Then, we fine-tune MentalBERT using our dataset to tailor it to the specific emotion-related student mental health during the COVID-19 pandemic. The model was trained for 5 epochs with a batch size of 16, learning rate of 2^{-5}, and a maximum sequence length of 128 tokens. The fine-tuning process adjusts the pre-trained model's parameters to minimize the loss on the training data using the true emotion labels in our dataset.

3.2.5 Zero-Shot GPT

GPT [30], a generative LLM, has revolutionized NLP and artificial intelligence since ChatGPT was introduced in 2022. In this study, we utilize OpenAI's GPT-3.5 [27] to generate emotion predictions for each response, as a baseline method. Given its ability to understand and generate human-like text, GPT-3.5 can be prompted with the students' responses and asked whether the input paragraph expresses specific emotions. The predictions are then mapped to the binary labels for further evaluation. The details of GPT's prompts are listed in Appendix B. We do not input the human labeling into the GPT prompt; thus, the method can be considered a zero-shot one. The experiment was conducted using GPT-3.5-turbo, the July 2024 version.

In Table 2, we summarize the training complexity of the five methods compared in this paper.

Table 2. The model complexity comparison of the five methods we compare in this study.

Model	Complexity
Lexicon	Build a customized dictionary for each emotion, usually including around 10 words in our application.
BoW	Convert each instance to a vector with maximum length 1000, and train a linear SVM (parameters $<$ 1k).
TF-IDF	Convert each instance to a vector with maximum length 1000, and train a logistic regression (parameters $<$ 1k).
Mental-BERT	Pretrained 110M parameters, fine-tuning for each emotion.
GPT-3.5	Pretrained 20B parameters, training is not needed.

3.3 Comparison Framework

3.3.1 Performance Criterion

To compare the performance of the different models, we use several standard evaluation metrics: True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), Accuracy, Precision, Recall, F1 Score, Receiver Operating Characteristic (ROC) Curve, and Area Under the ROC Curve (AUC). For a response r, we set $y_r(\mathbf{E}_j) = 1$ if human labeling considers that it includes the emotion $\mathbf{E}_j$, otherwise $y_r(\mathbf{E}_j) = 0$. Then, for an NLP method, its prediction result is $\hat{y}_r(\mathbf{E}_j)$. The first four metrics are calculated as:

$$\begin{aligned}
\mathrm{TP}_{\mathbf{E}_j} &= \sum_{r=1}^{R} \mathbf{1}\Big\{y_r(\mathbf{E}_j) = \hat{y}_r(\mathbf{E}_j) = 1\Big\} \\
\mathrm{FP}_{\mathbf{E}_j} &= \sum_{r=1}^{R} \mathbf{1}\Big\{y_r(\mathbf{E}_j) \neq 1 \text{ and } \hat{y}_r(\mathbf{E}_j) = 1\Big\} \\
\mathrm{FN}_{\mathbf{E}_j} &= \sum_{r=1}^{R} \mathbf{1}\Big\{y_r(\mathbf{E}_j) = 1 \text{ and } \hat{y}_r(\mathbf{E}_j) \neq 1\Big\} \\
\mathrm{TN}_{\mathbf{E}_j} &= \sum_{r=1}^{R} \mathbf{1}\Big\{y_r(\mathbf{E}_j) = \hat{y}_r(\mathbf{E}_j) \neq 1\Big\},
\end{aligned}$$

where $\mathbf{1}\{\cdot\}$ denotes an indicator function. Then, accuracy measures the proportion of correct predictions (both TP and TN) out of the total number of predictions, calculated as:

$$\mathrm{ACC}_{\mathbf{E}_j} = \frac{\mathrm{TP}_{\mathbf{E}_j} + \mathrm{TN}_{\mathbf{E}_j}}{\mathrm{TP}_{\mathbf{E}_j} + \mathrm{TN}_{\mathbf{E}_j} + \mathrm{FP}_{\mathbf{E}_j} + \mathrm{FN}_{\mathbf{E}_j}}. \tag{3.6}$$

Precision measures the proportion of true positive predictions out of all positive predictions (TP and FP), calculated as:

$$\mathrm{Precision}_{\mathbf{E}_j} = \frac{\mathrm{TP}_{\mathbf{E}_j}}{\mathrm{TP}_{\mathbf{E}_j} + \mathrm{FP}_{\mathbf{E}_j}}. \tag{3.7}$$

Recall measures the proportion of true positive predictions out of all actual positive cases (TP and FN), calculated as:

$$\mathrm{Recall}_{\mathbf{E}_j} = \frac{\mathrm{TP}_{\mathbf{E}_j}}{\mathrm{TP}_{\mathbf{E}_j} + \mathrm{FN}_{\mathbf{E}_j}}. \tag{3.8}$$

At last, the F1 Score is the harmonic mean of Precision and Recall, providing a single metric that balances both concerns, calculated as:

$$\mathrm{F1}_{\mathbf{E}_j} = 2 \times \frac{\mathrm{Precision}_{\mathbf{E}_j} \times \mathrm{Recall}_{\mathbf{E}_j}}{\mathrm{Precision}_{\mathbf{E}_j} + \mathrm{Recall}_{\mathbf{E}_j}}. \tag{3.9}$$

The Receiver Operating Characteristic (ROC) Curve is a graphical representation of a model's diagnostic ability. It plots the True Positive Rate (Recall) against the False Positive Rate (FPR) with different thresholds, where FPR is defined as:

$$\mathrm{FPR}_{\mathbf{E}_j} = \frac{\mathrm{FP}_{\mathbf{E}_j}}{\mathrm{FP}_{\mathbf{E}_j} + \mathrm{TN}_{\mathbf{E}_j}} \tag{3.10}$$

Then, we can calculate the Area Under the ROC Curve (AUC), which quantifies the model's overall ability to discriminate between positive and negative classes. A higher AUC indicates a better-performing model.

3.3.2 Comparison Steps

To give a comprehensive comparison among the NLP methods in the emotion detection from students' responses, our study includes the following three steps:

1. Compare the emotion detection performance of the five NLP methods. We use the accuracy and F1 scores as the criteria. For the four trainable methods, i.e., Lexicon, BoW, TF-IDF, and MentalBERT, we vary the size of the training data using 20%, 50%, and 80% of the entire dataset, to evaluate how the sample size affects the performance. We repeat the training/testing splits 100 times and report the average accuracy and F1 scores on the testing set. The zero-shot method, GPT-3.5, is used as the baseline.
2. Evaluate the stability and distinguishing ability of the NLP methods. We performed the 5-fold stratified cross-validation [45] 100 times with different data separations. In each stratified cross-validation, the data was split into five stratified folds, ensuring proportional representation of the positive/negative instances. The model is trained on four folds and tested on the remaining fold to evaluate performance. The process is repeated five times, once for each fold as the testing set. By doing so, we reduce variability caused by differences in the proportions of positive/negative instances between the training and testing datasets. The standard deviations of the accuracy and F1 scores are calculated to assess whether the model's performance is sensitive to data splitting. Then, we obtain the predicted probabilities for emotions and calculate the average AUC for each model across 100 stratified cross-validations to assess their distinguishing abilities.
3. Show the consistency of the detection results among the five NLP methods for different emotions. We conduct pairwise comparisons to determine whether the detection results of one approach are consistent with those of another. This analysis highlights the similarities and differences among the five NLP methods.

4. COMPARISON RESULTS AND DISCUSSION

4.1 Detection Performance Comparison

In this section, we examine the performance of four trainable methods – Lexicon, BoW, TF-IDF, and MentalBERT – as well as the zero-shot method, GPT. For the four trainable methods, we consider training data splits of 80%, 50%, and 20% to evaluate the impact of training sample size. We choose accuracy and F1 scores from the testing data, as defined in Section 3, as the performance criteria. For each percentage, we repeat the training/testing splits 100 times, and report the average performance criteria. The original results are presented in the tables in Appendix C.1.

To compare the performance of the five methods, we show the boxplots of their accuracy (upper panel) and F1-scores (lower panel) with various training percentages in Figure 3. We observe that the training percentage has a limited effect on the Lexicon's performance. The median F1 score of the Lexicon with a 20% training split is even slightly higher than those with 80% and 50% training percentages. The performance of BoW and TF-IDF shows modest impacts of the training percentage, and their overall trends are similar to each other. At last, the performance of MentalBERT, especially its F1 score, is significantly impacted by the training percentage. The median F1 score of MentalBERT with 80% is the highest among all NLP methods; however, its median F1 score with 20% is the lowest. A possible reason is that 20% of the training data, comprising only 80 instances, is insufficient to fine-tune the 110M parameters in the MentalBERT model. The zero-shot GPT demonstrates modest performance, similar to BoW and TF-IDF with a 50% training percentage. However, its F1 scores surpass BoW, TF-IDF, and MentalBERT with a 20% training percentage. Overall, the Lexicon method is relatively resistant to decreases in the training dataset, while MentalBERT is the most sensitive. Meanwhile, the performance changes in BoW and TF-IDF with varying training set sizes are moderate.

We examine the performance of those methods for individual emotions. We select three emotions: *Depression*, *Lack of Motivation*, and *Miscellaneous*, which present various levels of detection performance. The emotion with good performance, *Depression*, can be identified by keywords such as "depression" and "depressed" from the responses. The emotion associated with poor performance, *Miscellaneous*, is ambiguous and has only a small positive sample size. The moderate one, *Lack of Motivation*, has a clear definition but requires a comprehensive understanding of the responses to detect it. The accuracies and F1 scores of the three typical emotions of NLP methods with various training percentages are shown in Figure 4. We also highlight the values of the three typical emotions in Figure 3.

For *Depression*, the four trainable NLP methods with 80% training percentages achieve high performance with an accuracy of $0.9 \sim 1.0$ and F1 scores of $0.8 \sim 1.0$. Moreover, Lexicon, BoW, and MentalBERT all achieve significant improvements in both accuracy and F1 score compared to GPT, when the training percentage exceeds 50%, whereas TF-IDF yields a minor improvement. However, with a 20% training percentage, the performance of TF-IDF and MentalBERT drops dramatically, while the performance of Lexicon and BoW remains stable.

For the emotion with moderate detection difficulty, *Lack of Motivation*, the four trainable methods, with an 80% training percentage, can achieve higher or similar accuracy or F1 scores compared to the zero-shot GPT. However, the increase between the trainable methods and the GPT is smaller than that for the *Depression* case. The F1 score of the MentalBERT is superior to that of other methods,

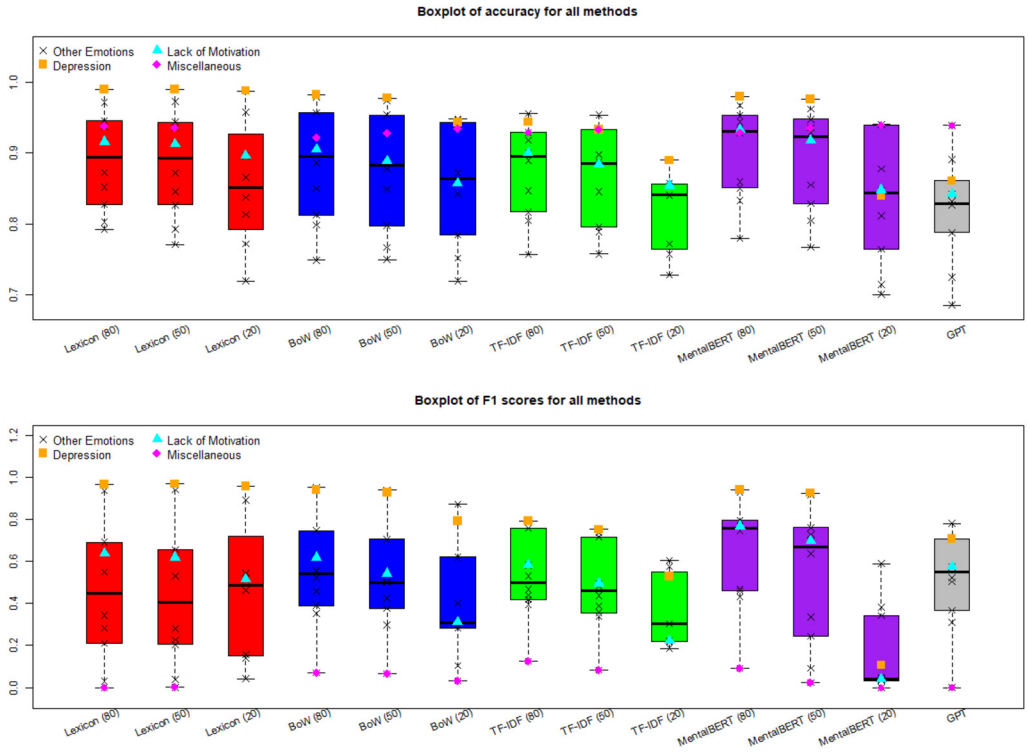

Figure 3: The boxplots of accuracy and F1 scores from the testing data among 10 emotions for the four trainable methods with 80%, 50%, and 20% training percentages and the zero-shot GPT. The values of the three typical emotions, *Depression*, *Lack of Motivation*, and *Miscellaneous*, are also highlighted.

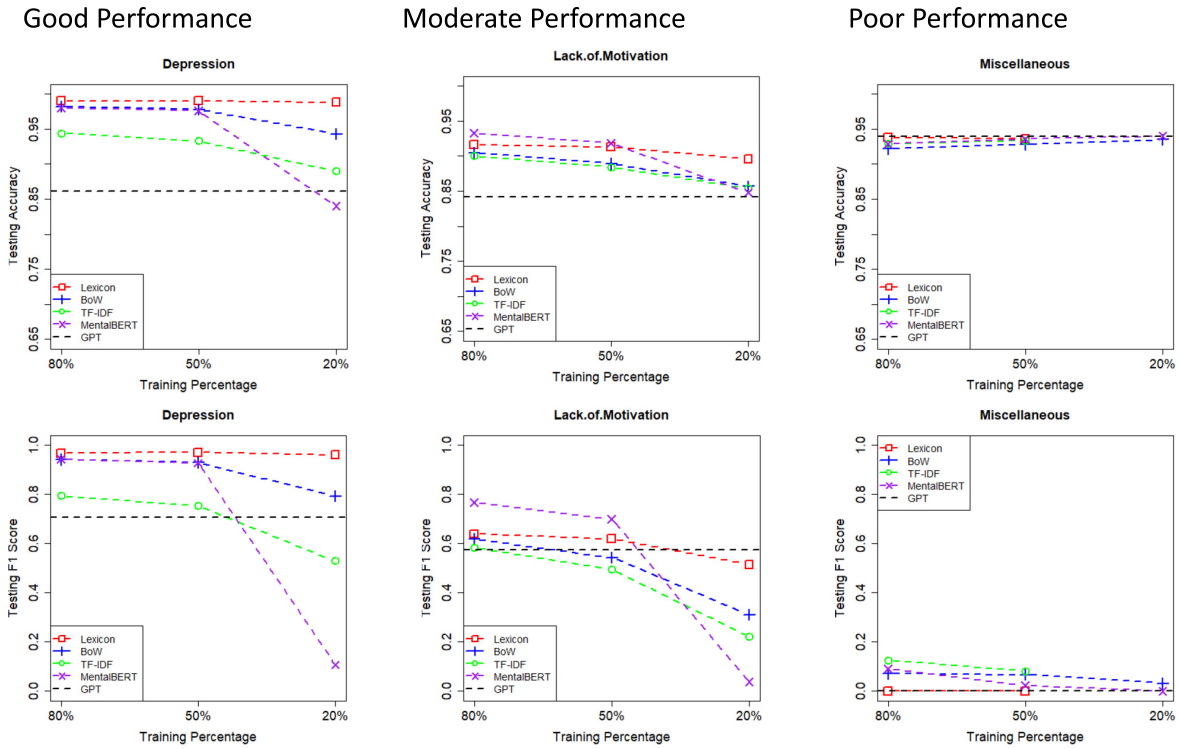

Figure 4: The accuracy and F1 score of the four trainable methods with 80%, 50%, and 20% training data split, and the zero-shot GPT for three typical emotions spanning different levels of detection performance.

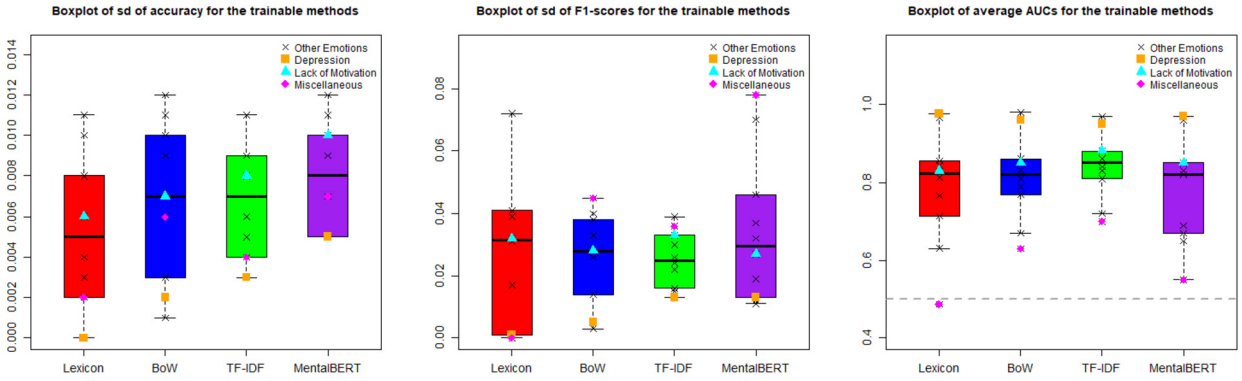

Figure 5: The boxplots of stability and distinguishing ability measurements among 10 emotions of the four trainable methods using 100 repetitions of 5-fold cross-validation. The values of the three typical emotions, *Depression*, *Lack of Motivation*, and *Miscellaneous*, are also highlighted.

demonstrating its capability to understand the context of the responses. When the training percentage reaches 50% and 20%, the F1 scores of most trainable methods decrease to levels below or similar to those of GPT. This suggests that we require a sufficiently large dataset to train the model for this emotion.

For the challenging emotion, *Miscellaneous*, the four trainable methods, along with GPT, achieve high accuracy of around $0.90 \sim 0.95$ and low F1-scores of around $0 \sim 0.1$. The results show that there is a very small number of positive data points for the *Miscellaneous* emotion, and the NLP methods will predict all sentences as negative. Due to this issue, the training percentages cannot help the performance of the four trainable results. Moreover, the zero-shot GPT also fails to recognize this ambiguous definition with a near-zero F1 score.

4.2 Model Stability and Distinguishing Ability

The next step is to evaluate the stability and distinguishing ability of the four trainable methods: Lexicon, BoW, TF-IDF, and MentalBERT. We find the mean and standard deviation of the accuracies and F1 scores from 100 repetitions of stratified 5-fold cross-validation. The means are similar to those with an 80% training percentage in Section 4.1, as the models using 5-fold cross-validation were also fitted from 80% of the data. The standard deviation tells us how the performance measurements change depending on which 80% of the data they are trained with, showing the stability of the methods. Then, we calculate the average AUC of the 100 repetitions, which indicates the model's ability to separate each emotion. The AUC is close to 1 when a method is capable of identifying all instances of an emotion with very few false positives. At the same time, an AUC of 0.5 means the method distinguishes an emotion no better than a random guess. The original results are presented in the tables in Appendix C.2.

The left and middle plots in Figure 5 show the boxplots of the standard deviations of the accuracy and F1 scores, and the right plot illustrates the boxplots of average AUCs among 10 emotions for the four trainable methods. There is no result for GPT, as it does not require a training process and cannot produce a probability of a positive detection. For the accuracy, it is clear that Lexicon has the smallest standard deviations, while MentalBERT has the largest. For the standard deviations of F1 scores, the trend is less obvious. However, the lower bound of Lexicon's boxplot is lower than that of the other three trainable methods, showing that it can achieve the highest stability for some emotions. Those plots show that Lexicon excels in the stability measurements, while MentalBERT's performance is sensitive to the part of the data with which it is trained. In the right plot, the upper bounds of the boxplots are all close to 1.0, but their lower bounds are different. Among the four methods, TF-IDF shows the highest average AUCs, and Lexicon and MentalBERT have the lowest values. Thus, when using AUC as the criterion, TF-IDF shows the highest distinguishing ability for some challenging emotions.

We show the stability and distinguishing ability measurements for the three typical emotions identified in Section 4.1. Their values are also highlighted in Figure 5 with special legends. For the emotion with good detection performance, *Depression*, the standard deviations of accuracy and F1 score are close to 0 for Lexicon, and slightly increase from BoW and TF-IDF to MentalBERT. The AUCs of all four methods are almost 1, showing excellent distinguishing ability. For the moderate one, *Lack of Motivation*, the standard deviation of accuracy again increases from Lexicon, BoW, TF-IDF, to MentalBERT. However, MentalBERT achieves the smallest standard deviation of F1 scores. One possible reason is that MentalBERT is better able to understand the concept of *Lack of Motivation*, thereby improving its stability. Their AUCs are around 0.75 and 0.85, and TF-IDF achieves the highest AUC.

For the emotion with poor performance, *Miscellaneous*, the standard deviations of Lexicon are still low despite its overall poor averages shown in Section 4.1. MentalBERT

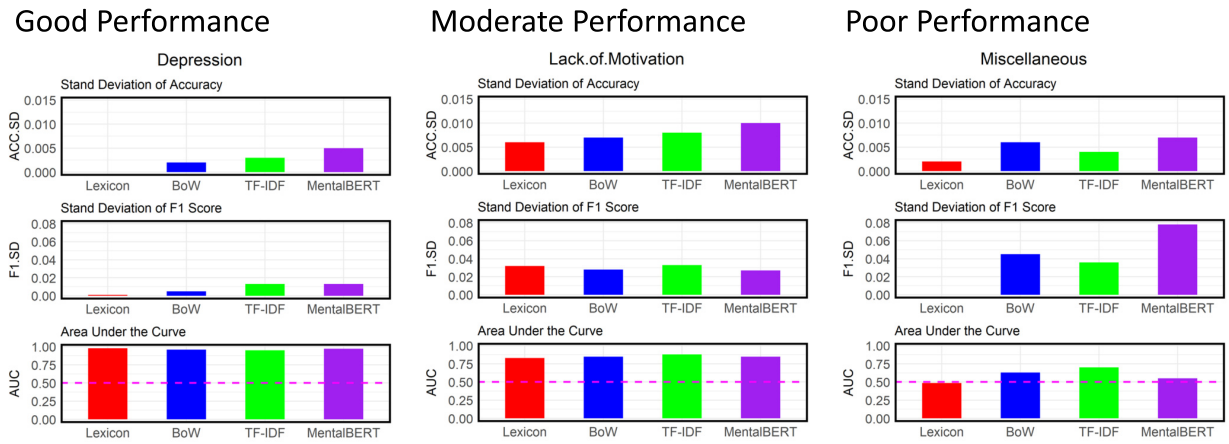

Figure 6: The stability and distinguishing ability measurements of the four trainable methods using 5-fold cross-validation for three typical emotions spanning different levels of detection performance.

again has the largest standard deviation in its accuracy and F1 scores. The AUCs are around 0.5 and 0.6, indicating the model's prediction is slightly better than the random guess. Among the four methods, TF-IDF has the highest AUC, while Lexicon and MentalBERT have the lowest. We can conclude that for both easy and challenging emotions, Lexicon exhibits the highest stability, while MentalBERT shows the lowest. For the moderate one, MentalBERT's stability becomes better. For moderate and challenging emotions, TF-IDF's distinguishing ability outperforms others when using AUC as the criterion.

4.3 Prediction Consistency

In this section, we aim to compare the consistency between the detections of different NLP methods, which indicates whether they yield identical predictions for each student's response. To handle emotions with a very small number of positive samples, we choose the Jaccard index, also known as the Jaccard similarity, to measure consistency. The Jaccard index between the two methods for a certain emotion can be calculated as:

$$\text{Jac}(\text{Meth}_1, \text{Meth}_2) = \frac{\sum_{r=1}^{R} I(\hat{y}_r^{\text{Meth}_1} = 1 \text{ and } \hat{y}_r^{\text{Meth}_2} = 1)}{\sum_{r=1}^{R} I(\hat{y}_r^{\text{Meth}_1} = 1 \text{ or } \hat{y}_r^{\text{Meth}_2} = 1)},$$

where $I(\cdot)$ is the indicator function, and $\hat{y}_r^{\text{Meth}_1}$ and $\hat{y}_r^{\text{Meth}_2}$ denote whether the two methods predict is that emotion expressed in response r. The prediction results are based on the first repetition of the 100 5-fold cross-validation processes in Section 4.2. We calculate the Jaccard indices from the 15 pair-wise comparisons between five NLP methods and the true labels, and the results of the ten emotions are listed in Appendix C.3. Figure 7 presents the average pair-wise consistencies over the ten emotions. MentalBERT vs. true labels (0.5) and BoW vs. TF-IDF (0.48) have the two highest average Jaccard indices. The first pair demonstrates the capability of trainable Transformer models to predict

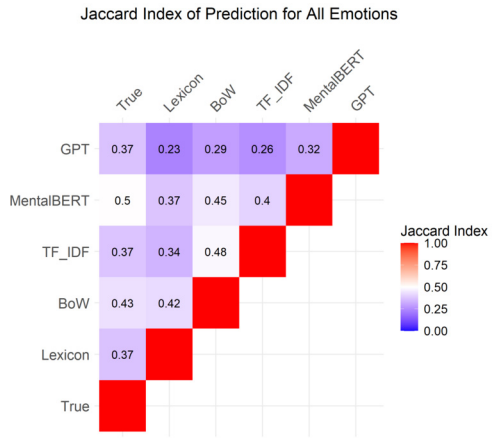

Figure 7: The average Jaccard indices among 10 emotions between five NLP methods and the true labels.

results close to the true labels. Meanwhile, the second pair is possibly caused by the similar mechanisms of the two methods, which first convert the responses to vectors and then train a machine learning classifier. Then, we examine the leftmost column, which shows the consistency of the five NLP methods compared to the true label. MentalBERT has the highest score (0.5), followed by BoW (0.43), and Lexicon, TF-IDF, and GPT have the lowest scores (0.37).

Finally, we present the Jaccard indices for the three typical emotions. For the good one, *Depression*, almost all pair-wise consistencies are above 0.5, while the consistencies between the true labels, Lexicon, BoW, and MentalBERT are higher than 0.75. However, the TF-IDF and GPT generate relatively inconsistent predictions, while their Jaccard index is lower than 0.5, which is consistent with the results in Figure 4 where these two methods show lower accuracies and F1 scores. For the emotion with moderate detection performance, *Lack of Motivation*, we find that MentalBERT

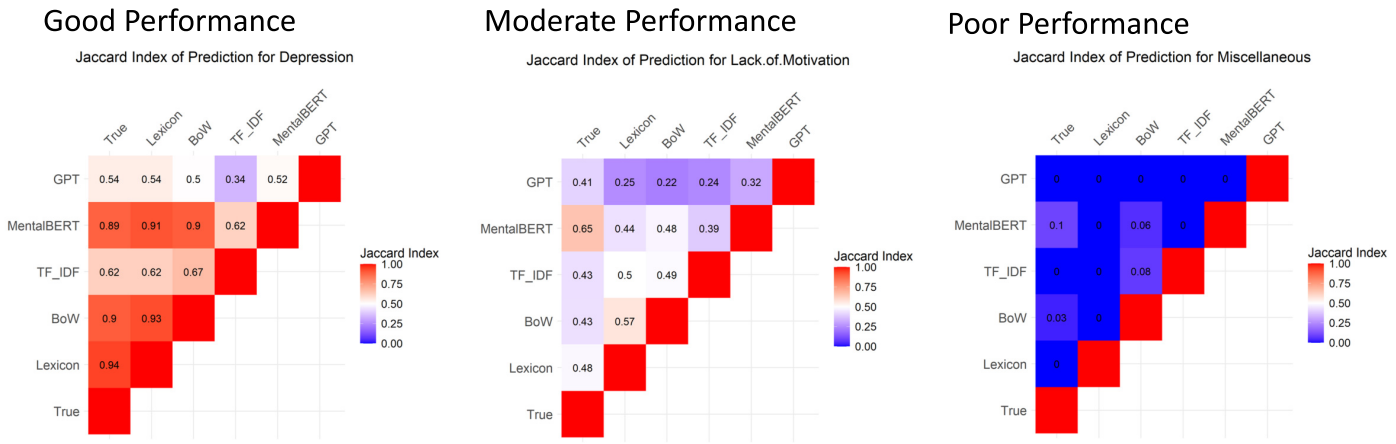

Figure 8: The Jaccard indices between five NLP methods and the true labels for three typical emotions spanning different levels of detection performance.

achieves the highest consistency with the true labels, which is around 0.6, while the other Jaccard indices range from 0.22 to 0.57. This result highlights the capabilities of trainable LLMs. For the emotion with poor performance, *Miscellaneous*, every pairwise comparison is below 0.1, indicating that none of the NLP methods, whether trainable or not, can capture the ambiguous concept of *Miscellaneous*.

4.4 Discussion

Performance and Complexity Trade-off: The NLP methods investigated in this study exhibit dramatically different levels of complexity. The Lexicon-based method requires learning scores for only dozens of keywords, while MentalBERT must fine-tune over 100 million parameters in its Transformer architecture [17]. These complexity differences translate into distinct performance patterns. The Lexicon-based method achieves stable performance across training sample sizes ranging from 20% to 80%, demonstrating the lowest standard deviations in both accuracy and F1 scores. However, its peak performance cannot compete with more sophisticated methods. In contrast, fine-tuned MentalBERT demonstrates the highest accuracy and F1 scores with 80% training data and achieves the best Jaccard Index agreement with true labels. Yet its performance drops dramatically with limited training data (20%), and it exhibits the highest variability in repeated cross-validation experiments.

BoW and TF-IDF methods, both of which convert responses into numeric vectors before training traditional machine learning models, demonstrate balanced performance, complexity, and stability. Their methodological similarity is reflected in their relatively high pairwise Jaccard index. While TF-IDF achieves better AUC scores than BoW, indicating good distinguishing ability for challenging emotions, its performance declines more rapidly with reduced training data.

The zero-shot GPT method, implemented through OpenAI's API, requires no training process and achieves performance comparable to BoW and TF-IDF models trained on 50% of the data, though it underperforms compared to MentalBERT trained on 80% of the data. Notably, Lossio-Venture et al. [19] found that zero-shot ChatGPT outperformed fine-tuned Transformers in sentiment analysis for COVID-19 survey data. This apparent discrepancy likely stems from task complexity differences: sentiment analysis predicts general positive/negative sentiment, a more universal task that ChatGPT's vast training data can handle effectively. In contrast, our fine-grained emotion detection requires distinguishing among ten mental health-related emotions, many specifically defined by our research team. Without access to labeled training data, the zero-shot GPT model cannot accurately detect these domain-specific emotional categories.

Method Selection Suggestions: Based on our performance observations, we offer the following guidelines for NLP method selection. The Lexicon-based method proves particularly effective for detecting well-defined emotions with clear linguistic indicators, especially when training data is limited or stability is prioritized over peak performance. Fine-tuned MentalBERT is most suitable for detecting contextually complex emotions when sufficient training data ($>$ 50% of available samples) and computational resources are available. Traditional machine learning methods, such as BoW and TF-IDF, provide effective predictions when labeled data or computational resources are insufficient for Transformer models. Finally, zero-shot GPT can generate quick assessments when no training data is available, though performance will be limited for domain-specific emotions.

Uncertainty in Mental Health Studies: When deploying NLP methods for emotion detection in mental health research, their inherent uncertainty must be carefully

examined, as detection results can have serious consequences for both research conclusions and potential interventions. Uncertainty arises from multiple sources, beginning with the training dataset itself. As demonstrated by our experiments, most NLP methods show performance sensitivity to training size variations. Additionally, repeated cross-validation reveals that resampling the training dataset while maintaining the same sample size yields variable predictions, particularly for high-complexity models like fine-tuned MentalBERT. Therefore, sensitivity analysis for training sample size and cross-validation repetitions is essential for evaluating model stability.

A second source of uncertainty stems from the binarization of predicted probabilities. All four trainable methods output probabilities indicating the likelihood of emotion presence, with positive detection determined by a 0.5 threshold. However, a response with 0.99 prediction probability represents a different uncertainty level compared to one with 0.51 probability. In this study, we employed AUC to evaluate the distinguishing ability of NLP models based on predicted probabilities, and ROC curves can provide additional insights into model uncertainty characteristics.

Emotion Labeling Impact: A crucial finding is that the target emotion significantly impacts NLP method performance. Figures 3 and 5 demonstrate that accuracy and stability measurements for the same method vary substantially across the ten emotions studied. Training sample size sensitivity also depends on the specific emotions being detected. Consequently, method comparisons yield different conclusions for different emotion types, as illustrated by our analysis of *Depression*, *Lack of Motivation*, and *Miscellaneous* categories. Such performance disparities across emotions can lead to inconsistent findings in studies relying on NLP detection results. While traditional mental health studies, such as [1], design emotion categories based on domain knowledge, the increasing role of NLP methods in data analysis necessitates careful selection and design of emotion categories to ensure both performance and stability of automated detection algorithms. For instance, *Miscellaneous*, a convenient category for human labelers, leads to poor performance for all NLP methods. Such categories should be avoided when incorporating NLP methods for data analysis.

We note that all four trainable methods can improve the model's consistency in producing predictions that are similar to those of human labelers. Thus, the zero-shot GPT-3.5 performs relatively poorly compared to other methods when we use human labels as the "golden standard". However, as pre-trained LLMs become more powerful, their predictions could be more valuable when the quality of human labels cannot be ensured. Collaboration between LLMs and human experts can be beneficial for mental health research.

Multi-label Classification: There are machine learning research related to multi-label classification [3, 40]. They found that by adopting specialized methods, such as problem transformation and algorithm selection, we can capture the inner structure of the labels to improve accuracy. In Figure 2, there are weak associations between our ten emotions. For example, *No/Positive Effects* have $-0.20 \sim -0.03$ correlations with other 9 emotions. Employing multi-label classification to improve the efficiency of emotion detection would be an interesting future study.

5. CONCLUSION

This paper presents a comprehensive comparative study of NLP methods for detecting fine-grained emotions in college student responses regarding their mental health during the COVID-19 pandemic. We evaluated five distinct approaches: Lexicon, BoW, TF-IDF, fine-tuned MentalBERT, and zero-shot GPT, examining their performance, training sample size sensitivity, stability, distinguishing ability, and inter-method consistency. Our experimental results reveal performance-complexity trade-offs among NLP methods and provide evidence-based guidelines for method selection. We demonstrate the critical importance of recognizing uncertainty inherent in NLP detections and emphasize the need for careful emotion category design to ensure detection quality. Our insights bridge a critical gap in NLP analysis between data science and mental health studies, utilizing survey data with various applications.

This work establishes a foundation for future NLP development in mental health survey research through several promising directions. First, hybrid or mixture-of-experts frameworks could be designed to balance performance-complexity trade-offs by selecting appropriate models based on emotion type and available training sample size, thereby providing stable detection results across diverse conditions. Second, uncertainty-aware algorithms could be developed based on our analytical framework, incorporating prediction probabilities, cross-validation standard deviations, and inter-method consistency scores to generate uncertainty estimates. In mental health applications, such systems could restrict automated decisions to low-uncertainty cases while flagging high-uncertainty responses for human review. Finally, our findings highlight the need for developing emotion categories that balance mental-health insights with computational detectability, potentially through collaborative efforts between domain experts and NLP researchers.

APPENDIX A. LABELING CRITERIA FOR TEN EMOTIONS

Table 3 gives an example for each emotion in our labeling process. Most of those emotions are self-explanatory, except the last three. We list our labeling criterion here:

- *Negative Feelings* include fear, grief or sorrow, sadness, hopelessness, lack of purpose or control;
- *All Stress* includes stress or worry related to the academy, financial and job, health, or general reasons;

Table 3. Examples of the 10 *emotions in manual labeling in our dataset.*

Emotions affecting mental health	**Sample of positive responses**
Isolation	"I'm extroverted and the isolation is very taxing."
Depression	"I am deeply depressed from watching the world collapse and feeling so helpless and useless."
Anxiety	"I had general anxiety even before the pandemic, so this has just been adding to it."
Issues With Home Life	"Being away from campus has taken my independence away which has taken a toll on me."
Lack of Routine	"I experience a lack of routine or structure due to the coronavirus."
Lack of Motivation	"The lack of a concrete schedule from not attending classes and extracurricular activities has distorted the passage of time and days feel much longer than they should be."
No/Positive Effects	"Giving me time to really focus on myself and develop new hobbies."
Negative Feelings	"I constantly have panic attacks or moments of sadness."
All Stress	"I feel like I am constantly stressed about me or my family contracting it, and it worries me because I care about my health, but even more about their health."
Miscellaneous	"This situation has caused me to struggle even more with my recovery from an eating disorder."

Table 4. The accuracy and F1 score of each method with 80%/20% *training/testing splitting.*

	Isolation	*Depression*	*Anxiety*	*Issues With Home Life*	*Lack of Routine*
Lexicon					
Accuracy	0.793	0.990	0.972	0.852	0.946
F1	0.693	0.967	0.937	0.282	0.211
BoW					
Accuracy	0.799	0.982	0.979	0.850	0.957
F1	0.748	0.942	0.953	0.459	0.524
TF-IDF					
Accuracy	0.817	0.944	0.918	0.847	0.956
F1	0.758	0.794	0.793	0.441	0.467
MentalBERT					
Accuracy	0.833	0.980	0.967	0.860	0.953
F1	0.797	0.942	0.929	0.470	0.461

- *Miscellaneous* includes miscellaneous mental health issues, such as self-harm, substance abuse, trauma, eating disorders, sleep issues, etc.

APPENDIX B. PROMPTS FOR GPT-3.5

Using OpenAI's API, we first write the following prompt:

```
messages = [
      {"role": "system", "content": "You are categorizing
           the emotions and issues expressed by the student
           in their response. A response can possess
           multiple emotions/issues."},
      {"role": "user", "content": "Emotions and issues:
           Isolation/Loneliness, Depression, Anxiety,
           Negative Feelings, Lack of Motivation, All
           Stress (including academic, financial, health),
           Miscellaneous Mental Health Issues, Home Life
           Issues, Positive Effects, Lack of Routine."},
      {"role": "user", "content": "Student response: " +
           response}
   ]
```

where `response` is the students' responses r in their survey. We then send the message to GPT-3.5:

```
completion = client.chat.completions.create(
      model="gpt-3.5-turbo",
      messages=messages,
      temperature=0.7,
      max_tokens=250,
      )
```

When receiving the answer from GPT-3.5, we will check whether it includes each emotion $\mathbf{E}_j$. If so, the corresponding detection $\hat{y}_r(\mathbf{E}_j)$ will be set as 1; otherwise $\hat{y}_r(\mathbf{E}_j) = 0$.

APPENDIX C. ORIGINAL RESULTS

C.1 Accuracy and F1 Scores with Various Training Sample Sizes

For the four trainable NLP methods, their accuracy and F1 scores from testing dataset with various training sample sizes for the ten emotions are listed in Tables 4 and 5 (80%/20% training/testing), Tables 6 and 7 (50%/50% training/testing), Tables 8 and 9 (20%/80% training/testing). The accuracy and F1 scores of the zero-shot GPT are listed in Table 10. We repeat the splitting 100 times and report the average performance criteria.

Table 5. The accuracy and F1 score of each method with 80%/20% training/testing splitting.

	Lack of Motivation	*No/Positive Effects*	*Negative Feelings*	*Miscellaneous*	*All Stress*
Lexicon					
Accuracy	0.916	0.873	0.803	0.938	0.828
F1	0.639	0.030	0.344	0.000	0.551
BoW					
Accuracy	0.905	0.886	0.749	0.922	0.813
F1	0.619	0.353	0.391	0.071	0.557
TF-IDF					
Accuracy	0.900	0.890	0.757	0.929	0.805
F1	0.583	0.395	0.418	0.125	0.531
MentalBERT					
Accuracy	0.933	0.944	0.780	0.929	0.851
F1	0.768	0.749	0.431	0.091	0.770

Table 6. The accuracy and F1 score of each method with 50%/50% training/testing splitting.

	Isolation	*Depression*	*Anxiety*	*Issues With Home Life*	*Lack of Routine*
Lexicon					
Accuracy	0.771	0.990	0.973	0.846	0.944
F1	0.658	0.970	0.942	0.227	0.205
BoW					
Accuracy	0.767	0.978	0.974	0.849	0.954
F1	0.706	0.930	0.943	0.425	0.500
TF-IDF					
Accuracy	0.789	0.933	0.898	0.846	0.954
F1	0.717	0.753	0.737	0.389	0.439
MentalBERT					
Accuracy	0.805	0.976	0.962	0.855	0.948
F1	0.762	0.927	0.917	0.092	0.245

Table 7. The accuracy and F1 score of each method with 50%/50% training/testing splitting.

	Lack of Motivation	*No/Positive Effects*	*Negative Feelings*	*Miscellaneous*	*All Stress*
Lexicon					
Accuracy	0.913	0.872	0.793	0.936	0.827
F1	0.619	0.039	0.280	0.001	0.531
BoW					
Accuracy	0.889	0.878	0.750	0.928	0.798
F1	0.543	0.298	0.376	0.067	0.503
TF-IDF					
Accuracy	0.884	0.888	0.758	0.933	0.796
F1	0.496	0.339	0.357	0.082	0.483
MentalBERT					
Accuracy	0.918	0.929	0.767	0.936	0.829
F1	0.699	0.637	0.336	0.023	0.730

Table 8. The accuracy and F1 score of each method with 20%/80% training/testing splitting.

	Isolation	*Depression*	*Anxiety*	*Issues With Home Life*	*Lack of Routine*
Lexicon					
Accuracy	0.720	0.988	0.958	0.838	NA
F1	0.546	0.960	0.894	0.141	NA
BoW					
Accuracy	0.720	0.943	0.947	0.843	0.948
F1	0.621	0.793	0.874	0.302	0.283
TF-IDF					
Accuracy	0.728	0.890	0.858	0.841	NA
F1	0.607	0.529	0.576	0.187	NA
MentalBERT					
Accuracy	0.701	0.840	0.812	0.940	0.941
F1	0.591	0.107	0.342	0.035	0.000

Table 9. The accuracy and F1 score of each method with 20%/80% *training/testing splitting.*

	Lack of Motivation	*No/Positive Effects*	*Negative Feelings*	*Miscellaneous*	*All Stress*
Lexicon					
Accuracy	0.896	0.866	0.772	NA	0.814
F1	0.515	0.042	0.158	NA	0.463
BoW					
Accuracy	0.857	0.872	0.752	0.935	0.785
F1	0.311	0.104	0.299	0.031	0.401
TF-IDF					
Accuracy	0.854	NA	0.758	NA	0.772
F1	0.222	NA	0.216	NA	0.305
MentalBERT					
Accuracy	0.848	0.878	0.765	0.940	0.715
F1	0.038	0.047	0.032	0.000	0.383

Table 10. The accuracy and F1 score of zero-shot GPT method.

	Isolation	*Depression*	*Anxiety*	*Issues With Home Life*	*Lack of Routine*
GPT					
Accuracy	0.788	0.861	0.827	0.827	0.832
F1	0.782	0.708	0.717	0.523	0.367
	Lack of Motivation	*No/Positive Effects*	*Negative Feelings*	*Miscellaneous*	*All Stress*
GPT					
Accuracy	0.842	0.891	0.686	0.939	0.725
F1	0.575	0.505	0.31	0	0.577

Table 11. The standard deviations of accuracy and F1 scores obtained from 100 *repetitive five-fold cross-validation for four trainable NLP methods.*

5-fold Cross Validation	*Iso-lation*	*Depression*	*Anxie-ty*	*Issues With Home Life*	*Lack of Routine*	*Lack of Motivation*	*No/ Positive Effects*	*Nega-tive Feelings*	*Miscellane-ous*	*All Stress*
					Lexicon					
Accuracy	0.793 ± 0.010	0.990 ± 0.000	0.976 ± 0.000	0.852 ± 0.006	0.944 ± 0.003	0.915 ± 0.006	0.871 ± 0.004	0.802 ± 0.008	0.938 ± 0.002	0.831 ± 0.011
F1	0.696 ± 0.017	0.971 ± 0.001	0.947 ± 0.000	0.275 ± 0.041	0.202 ± 0.072	0.633 ± 0.032	0.037 ± 0.031	0.335 ± 0.039	0.000 ± 0.000	0.555 ± 0.041
					BoW					
Accuracy	0.797 ± 0.011	0.982 ± 0.002	0.977 ± 0.001	0.851 ± 0.009	0.958 ± 0.003	0.901 ± 0.007	0.889 ± 0.007	0.751 ± 0.012	0.922 ± 0.006	0.813 ± 0.010
F1	0.748 ± 0.014	0.944 ± 0.005	0.949 ± 0.003	0.468 ± 0.033	0.567 ± 0.038	0.618 ± 0.028	0.403 ± 0.040	0.398 ± 0.028	0.091 ± 0.045	0.562 ± 0.026
					TF-IDF					
Accuracy	0.818 ± 0.011	0.943 ± 0.003	0.916 ± 0.005	0.847 ± 0.008	0.957 ± 0.003	0.900 ± 0.008	0.891 ± 0.006	0.759 ± 0.011	0.930 ± 0.004	0.805 ± 0.009
F1	0.761 ± 0.016	0.802 ± 0.013	0.793 ± 0.015	0.458 ± 0.024	0.511 ± 0.030	0.600 ± 0.033	0.402 ± 0.039	0.428 ± 0.026	0.143 ± 0.036	0.541 ± 0.022
					MentalBERT					
Accuracy	0.822 ± 0.009	0.981 ± 0.005	0.967 ± 0.005	0.865 ± 0.009	0.954 ± 0.005	0.930 ± 0.010	0.942 ± 0.007	0.783 ± 0.011	0.934 ± 0.007	0.853 ± 0.012
F1	0.789 ± 0.012	0.943 ± 0.013	0.929 ± 0.011	0.503 ± 0.046	0.471 ± 0.070	0.765 ± 0.027	0.741 ± 0.032	0.456 ± 0.037	0.159 ± 0.078	0.776 ± 0.019

C.2 Stability and Distinguishing Ability

Table 11 shows the standard deviations of accuracy and F1 scores obtained from 100 repetitive five-fold cross-validation for four trainable NLP methods. Table 12 lists the average Area Under ROC Curves (AUC) obtained from the stratified cross-validation for four trainable NLP methods.

C.3 Pairwise Consistency

Pairwise consistency between five NLP methods and the true labels for each emotion is listed in Figure 9.

ACKNOWLEDGEMENTS

The authors thank Alexis West and Afton White from Virginia Commonwealth University for categorizing emotions to produce human labeling results.

Table 12. The average Area Under ROC Curves (AUC) obtained from the first five-fold cross-validation for four trainable NLP methods.

	Iso-lation	*Depre-ssion*	*Anxie-ty*	*Issues With Home Life*	*Lack of Routine*	*Lack of Motiva-tion*	*No/ Pos-itive Ef-fects*	*Nega-tive Feelings*	*Misce-llane-ous*	*All Stress*
Lexicon	0.855	0.977	0.965	0.843	0.767	0.829	0.631	0.714	0.487	0.813
BoW	0.860	0.960	0.980	0.810	0.830	0.850	0.770	0.670	0.630	0.790
TF-IDF	0.880	0.950	0.970	0.840	0.860	0.880	0.810	0.720	0.700	0.830
Mental-BERT	0.820	0.970	0.960	0.690	0.670	0.850	0.820	0.650	0.550	0.830

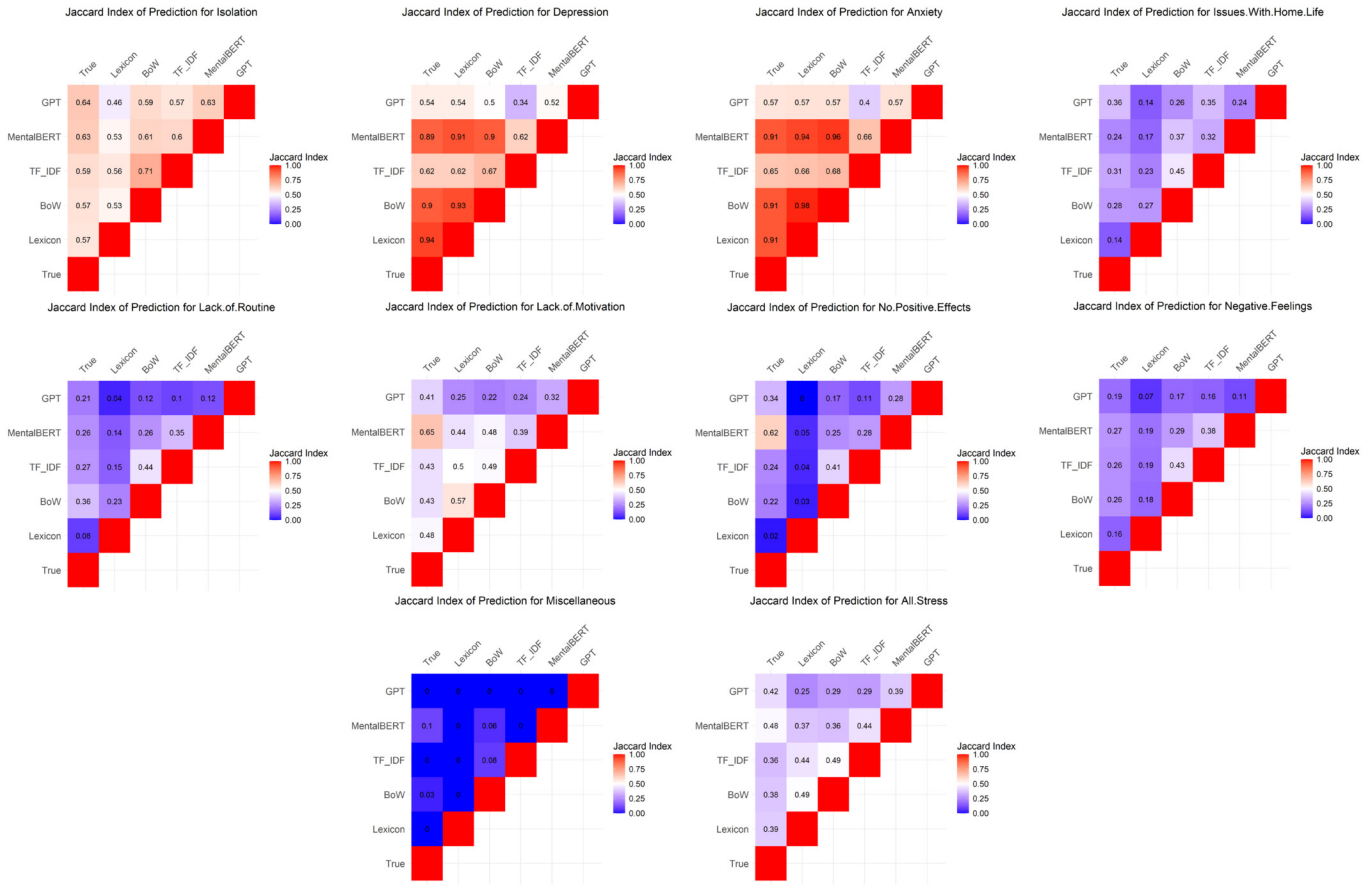

Figure 9: The pairwise comparison between five AI methods and the true label for the ten emotions.

FUNDING

This research was supported by NSF Research Experiences for Undergraduates (REU), grant number DMS1950015, and by the VCU College of Humanities and Sciences Catalyst.

Accepted 19 May 2026

REFERENCES

[1] Amona, E., West, A., White, A., Sahoo, I., Chan, D. M., Gandhi, P. and Qian, Y. (2025). Breakdown of COVID effects on students' mental health at the beginning of the pandemic. *PLOS Mental Health* **2**(6) 0000363.

[2] Barry, J. (2017). Sentiment Analysis of Online Reviews Using Bag-of-Words and LSTM Approaches. In *AICS* 272–274.

[3] Bogatinovski, J., Todorovski, L., Džeroski, S. and Kocev, D. (2022). Comprehensive comparative study of multi-label classification methods. *Expert Systems with Applications* **203** 117215.

[4] Boon-Itt, S., Skunkan, Y. et al. (2020). Public perception of the COVID-19 pandemic on Twitter: sentiment analysis and topic modeling study. *JMIR Public Health and Surveillance* **6**(4) 21978.

[5] Bouazizi, M. and Ohtsuki, T. (2019). Multi-class sentiment analysis on Twitter: Classification performance and challenges. *IEEE Access* **7** 46273–46284.

[6] Browning, M. H., Larson, L. R., Sharaievska, I., Rigolon,

A., McAnirlin, O., Mullenbach, L., Cloutier, S., Vu, T. M., Thomsen, J., Reigner, N. et al. (2021). Psychological impacts from COVID-19 among university students: Risk factors across seven states in the United States. *PloS One* **16**(1) 0245327.

[7] Chan, D. M., Broda, M. D., Winslow, J., Jones, Q., Luce, C., McGinnis, H. A., Tomlinson, C. A., Hamid, H. and Ma, J. (2022). The Effects of Prime Supporters within a College Student's Support Network. *Nonlinear Dynamics, Psychology & Life Sciences* **26**(4).

[8] Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* **16** 321–357.

[9] Copeland, W. E., McGinnis, E., Bai, Y., Adams, Z., Nardone, H., Devadanam, V., Rettew, J. and Hudziak, J. J. (2021). Impact of COVID-19 pandemic on college student mental health and wellness. *Journal of the American Academy of Child & Adolescent Psychiatry* **60**(1) 134–141.

[10] Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G. and Ravi, S. (2020). GoEmotions: A Dataset of Fine-Grained Emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* 4040–4054.

[11] Desmet, B. and Hoste, V. (2013). Emotion detection in suicide notes. *Expert Systems with Applications* **40**(16) 6351–6358. https://doi.org/10.1016/j.eswa.2013.05.050.

[12] Devlin, J., Chang, M. -W., Lee, K. and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 4171–4186.

[13] Floridi, L. and Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines* **30** 681–694.

[14] Guo, X., Zhang, G., Wang, S. and Chen, Q. (2020). Multi-way matching based fine-grained sentiment analysis for user reviews. *Neural Computing and Applications* **32**(12) 7729–7743.

[15] Hofmann, T., Schölkopf, B. and Smola, A. J. (2008). Kernel methods in machine learning. *The Annals of Statistics* **36**(3) 1171–1220. https://doi.org/10.1214/009053607000000677. MR2418654

[16] Jain, B., Goyal, G. and Sharma, M. (2024). Evaluating Emotional Detection & Classification Capabilities of GPT-2 & GPT-Neo Using Textual Data. In *2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* 12–18. https://doi.org/10.1109/Confluence60223.2024.10463396.

[17] Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P. and Cambria, E. (2022). Mentalbert: Publicly available pretrained language models for mental healthcare. In *Proceedings of the 13th Language Resources and Evaluation Conference* 7184–7190.

[18] Kim, H., Rackoff, G. N., Fitzsimmons-Craft, E. E., Shin, K. E., Zainal, N. H., Schwob, J. T., Eisenberg, D., Wilfley, D. E., Taylor, C. B. and Newman, M. G. (2022). College mental health before and during the COVID-19 pandemic: results from a nationwide survey. *Cognitive Therapy and Research* **46**(1) 1–10.

[19] Lossio-Ventura, J. A., Weger, R., Lee, A. Y., Guinee, E. P., Chung, J., Atlas, L., Linos, E. and Pereira, F. (2024). A comparison of ChatGPT and fine-tuned open pre-trained transformers (OPT) against widely used sentiment analysis tools: sentiment analysis of COVID-19 survey data. *JMIR Mental Health* **11** 50150.

[20] Mohammad, S. M. (2018). Word Affect Intensities. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan. https://aclanthology.org/L18-1027.

[21] Mohammad, S. M. and Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence* **29**(3) 436–465. https://doi.org/10.1111/j.1467-8640.2012.00460.x. MR3093841

[22] Mustafa, R. U., Ashraf, N., Ahmed, F. S., Ferzund, J., Shahzad, B. and Gelbukh, A. (2020). A Multiclass Depression Detection in Social Media Based on Sentiment Analysis. In *International Conference on Intelligent Systems Design and Applications* 879–889 Springer.

[23] Nadkarni, P. M., Ohno-Machado, L. and Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association* **18**(5) 544–551.

[24] Nardi, P. M. (2018) *Doing survey research: A guide to quantitative methods*. Routledge, an imprint of the Taylor & Francis Group.

[25] Naseem, U., Razzak, I., Khushi, M., Eklund, P. W. and Kim, J. (2021). COVIDSenti: A large-scale benchmark Twitter data set for COVID-19 sentiment analysis. *IEEE Transactions on Computational Social Systems* **8**(4) 1003–1015.

[26] Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N. and Mian, A. (2023). A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*.

[27] OpenAI (2023). GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

[28] Pröllochs, N., Feuerriegel, S. and Neumann, D. (2018). Statistical inferences for polarity identification in natural language. *PloS one* **13**(12) 0209323.

[29] Qader, W. A., Ameen, M. M. and Ahmed, B. I. (2019). An Overview of Bag of Words: Importance, Implementation, Applications, and Challenges. In *2019 International Engineering Conference (IEC)* 200–204. https://doi.org/10.1109/IEC47844.2019.8950616.

[30] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. et al. (2018). Improving language understanding by generative pre-training. Technical Report, OpenAI. https://cdn.openai.com/research-covers/language-unsupervised/.

[31] Rahman, S. S. M. M., Biplob, K. B. M. B., Rahman, M. H., Sarker, K. and Islam, T. (2020). An investigation and evaluation of N-Gram, TF-IDF and ensemble methods in sentiment classification. In *Cyber Security and Computer Science: Second EAI International Conference, ICONCS 2020, Dhaka, Bangladesh, February 15-16, 2020, Proceedings 2* 391–402. Springer.

[32] Ramos, J. et al. (2003). Using TF-IDF to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning* **242** 29–48. Citeseer.

[33] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)* **34**(1) 1–47.

[34] Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing* **14** 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88. MR2086398

[35] Son, C., Hegde, S., Smith, A., Wang, X. and Sasangohar, F. (2020). Effects of COVID-19 on college students' mental health in the United States: Interview survey study. *Journal of Medical Internet Research* **22**(9) 21279.

[36] Stoltzfus, J. C. (2011). Logistic regression: a brief primer. *Academic Emergency Medicine* **18**(10) 1099–1104.

[37] Sundaram, V., Ahmed, S., Muqtadeer, S. A. and Reddy, R. R. (2021). Emotion analysis in text using TF-IDF. In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* 292–297. IEEE.

[38] Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics* **37**(2) 267–307.

[39] Tang, T., Tang, X. and Yuan, T. (2020). Fine-tuning BERT for multi-label sentiment analysis in unbalanced code-switching text. *IEEE Access* **8** 193248–193256.

[40] Tsoumakas, G. and Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining* **3**(3) 1–13.

[41] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems* **30**.

[42] Wang, X., Hegde, S., Son, C., Keller, B., Smith, A. and

Sasangohar, F. (2020). Investigating mental health of US college students during the COVID-19 pandemic: Cross-sectional survey study. *Journal of Medical Internet Research* **22**(9) 22817.

[43] Weger, R., Lossio-Ventura, J. A., Rose-McCandlish, M., Shaw, J. S., Sinclair, S., Pereira, F., Chung, J. Y., Atlas, L. Y. et al. (2023). Trends in language use during the COVID-19 pandemic and relationship between language use and mental health: text analysis based on free responses from a longitudinal study. *JMIR Mental Health* **10**(1) 40899.

[44] Wright, L., Burton, A., McKinlay, A., Steptoe, A. and Fancourt, D. (2022). Public opinion about the UK government during COVID-19 and implications for public health: A topic modeling analysis of open-ended survey response data. *PloS One* **17**(4) 0264134.

[45] Zeng, X. and Martinez, T. R. (2000). Distribution-balanced stratified cross-validation for accuracy estimation. *Journal of Experimental & Theoretical Artificial Intelligence* **12**(1) 1–12.

Alexander Maret. Department of Mathematics, University of Pennsylvania, United States. E-mail address: maret@upenn.edu

Cade Dees. Department of Computer Science, University of Alabama, United States. E-mail address: cade.dees@yahoo.com

Yule Fu. Department of Mathematics, Duke University, United States. E-mail address: yule.fu@duke.edu

Yanjun Qian. Department of Statistical Sciences and Operations Research, Virginia Commonwealth University, United States. E-mail address: yqian3@vcu.edu

David Chan. Department of Mathematics and Applied Mathematics, Virginia Commonwealth University, United States. E-mail address: dmchan@vcu.edu

Punit Gandhi. Department of Mathematics and Applied Mathematics, Virginia Commonwealth University, United States. E-mail address: gandhipr@vcu.edu

Indranil Sahoo. Department of Statistical Sciences and Operations Research, Virginia Commonwealth University, United States. E-mail address: sahooi@vcu.edu