Understanding Student Sentiment on Mental Health Support in Colleges Using Large Language Models

Palak Sood, Chengyang He, Divyanshu Gupta, Yue Ning, Ping Wang
Department of Computer Science
Stevens Institute of Technology
Hoboken NJ, USA
{psood, che14, dgupta12, yue.ning, pwang44}@stevens.edu

***Abstract*—Mental health support in colleges is vital in educating students by offering counseling services and organizing supportive events. However, evaluating its effectiveness faces challenges like data collection difficulties and lack of standardized metrics, limiting research scope. Student feedback is crucial for evaluation but often relies on qualitative analysis without systematic investigation using advanced machine learning methods. This paper uses public Student Voice Survey data to analyze student sentiments on mental health support with large language models (LLMs). We created a sentiment analysis dataset, SMILE-College, with human-machine collaboration. The investigation of both traditional machine learning methods and state-of-the-art LLMs showed the best performance of GPT-3.5 and BERT on this new dataset. The analysis highlights challenges in accurately predicting response sentiments and offers practical insights on how LLMs can enhance mental health-related research and improve college mental health services. This data-driven approach will facilitate efficient and informed mental health support evaluation, management, and decision-making.**

***Index Terms*—Sentiment analysis, Mental health, Sentiment annotation, Text mining, Large language models**

I. Introduction

Mental health has become a paramount concern within the student community, increasingly recognized as essential to both their overall well-being and academic success [1], [2]. A 2020 report by the National Institute of Mental Health highlights that mental illness prevalence is highest among those under 25 years, including 67% of college students [3]. Universities play a crucial role by offering counseling services and organizing events to support students' emotional well-being. However, evaluating mental health support in colleges faces challenges such as data collection difficulties, lack of standardized metrics, insufficient funding, and limited inter-institutional collaboration [4], [5]. These issues restrict research scope, with most studies [6]–[8] focusing on student mental health status rather than the effectiveness of support services.

Student feedback is vital for assessing university mental health services. Surveys like the Healthy Minds Study [9] and the American College Health Association Health Assessment [10] gather insights into students' mental health and service utilization. Universities can use this feedback to improve their initiatives. Recent studies have explored student perspectives on mental health support [11], [12], but key challenges remain. These include reliance on qualitative analysis from limited feedback, lack of comprehensive quantitative evaluations, and the absence of utilizing advanced machine learning methods to analyze sentiments. Additionally, existing datasets do not support developing machine learning models for this purpose.

This paper aims to utilize the Student Voice Survey (SVS) data by College Pulse [13] to create a sentiment analysis dataset and explore the potential of large language models (LLMs) for predicting students' sentiments. Specifically, we utilize the students' narrative feedback in SVS data about their feedback for the advantages and disadvantages of mental health support in their colleges. To create the dataset for sentiment analysis, we first explore the spectrum of students' sentiments by leveraging the power of LLMs, considering the standard three categories of sentiment labels (including "Satisfied", "Dissatisfied", and "Neutral"), and designing a task-specific coarse prompt. This investigation motivates us to adopt a more detailed analysis by introducing a new sentiment category "Mixed". With the more nuanced set of sentiment categories, we collect the sentiment labels of students' responses in SVS data with human annotation, validation, and collaborative discussion.

The enriched SVS data, named **S**enti**M**ent analys**I**s of students' menta**L** h**E**alth support in **College**s (SMILE-College), includes 793 records with sentiment labels and is publicly available online[1]. Representative examples for each category are shown in Table I. We aim to investigate three tasks: (1) *Sentiment prediction*: Automatically predicting sentiment labels by designing task-specific prompts for LLMs with fine-grained sentiment categories. (2) *Prediction error analysis*: Analyzing LLM prediction errors across sentiment categories. (3) *Support limitation identification*: Using LLMs, embedding learning, and clustering techniques to identify key limitations of mental health support based on "Dissatisfied" responses. To the best of our knowledge, this is the first work to comprehensively evaluate student mental health support in colleges from students' perspective. This data-driven study enables the automatic prediction of students' perceptions of mental health support with advanced LLMs, providing quantitative and qualitative assessment.

[1] https://github.com/LEAF-Lab-Stevens/SMILE-College

Table I: Representative examples for each sentiment category.

Label	**Students' Survey Response**
Satisfied	I honestly think all of it is amazing so far, I visit the therapists and nurses a lot right now and it's all been covered by tuition and fees. everyone is super friendly and I always leave feeling like I had everything taken care of
Dissatisfied	I only know of one mental health employee but not know how to reach them or what to do . the therapy they provide is also catholic based which I am not
Mixed	my college works well in communicating about the various mental health resources on campus. more attention is needed to expand the mental health department in its diversity.
Neutral	I haven't personally used any of the services, so I feel as though I am not qualified to answer this question.

In summary, our study makes the following **key contributions**.

- Created the first sentiment analysis dataset of student mental health support in colleges by annotating sentiment labels with human-machine collaboration.
- Investigated several state-of-the-art LLMs on the SMILE-College data for three important sentiment prediction-related tasks with a fine-grained prompt.
- Experimental results highlight the better performance of GPT-3.5 and BERT on this specific task and underscore the challenges in accurately predicting the response sentiments.
- Identified key limitations of mental health support for potential improvement in colleges by leveraging the power of LLMs, embedding learning, and clustering techniques.

II. Related Work

The significance of mental health within student communities has escalated, underscoring the essential role of support services. Existing research about mental health in colleges mostly focuses on investigating students' mental health status [6]–[8]. Research on evaluating mental health support in colleges is limited due to various challenges. Various surveys, such as the Student Voice Survey and the Healthy Minds Study [9], are designed to gain students' insights on mental health services for assessing these services. The American College Health Association also conducted a survey to collect students' perceptions, behaviors, and habits [10], [14]–[16]. There are some recent works about student perspectives on improving mental health support services in universities or systematically reviewing the students' use of mental health services in universities [11], [12]. However, there are still several key challenges that have not been addressed, such as qualitative analysis on limited feedback, lack of comprehensive quantitative evaluation, and the absence of utilizing advanced machine learning methods for the evaluation.

Sentiment analysis is the computational study of opinions, attitudes, and emotions expressed in narrative texts [17]. Deep learning models, including recurrent neural networks and transformer-based models, have been successfully employed in sentiment analysis [18]–[21], while lexicon- and rule-based methods relying on sentiment dictionaries have also been widely used [22]. Sentiment analysis has also been applied to social media data for the detection of signs of depression and suicidal ideation, as demonstrated by Shen et al. [23]. Recently, the pre-trained and large language models gained significant attention in the field of sentiment analysis [24]. Sentiment analysis has been widely applied to various applications, such as social media [25] and customer feedback [26], and demonstrated its effectiveness. It has also been used in various mental health prediction and analysis tasks [27]. Most studies focus on analyzing individuals' mental health by examining their emotional sentiments, leveraging sentiment analysis to understand mental health states or detect early signs of mental health disorders [28], [29].

However, to the best of our knowledge, no prior work has explored evaluating students' perceptions of mental health support in colleges using sentiment analysis. The lack of such research leaves a critical gap in understanding how students perceive and engage with the mental health support resources available to them in colleges, which is essential for developing a more effective, student-centered mental health support system for each college. This paper fills this gap by introducing the first sentiment analysis dataset with a specific focus on student perceptions of mental health support in college settings. By creating and analyzing the dataset with LLMs, this work provides a foundation for data-driven evaluation and decision-making and offers insights into students' satisfaction and concerns based on their experiences with available mental health services. More importantly, the dataset will not only support related research into the sentiment analysis associated with mental health support but also has the potential to identify support limitations for actionable strategies to tailor mental health support to students' real needs.

III. The SMILE-College Dataset

To the best of our knowledge, no existing dataset has been specifically developed for sentiment analysis on mental health support in colleges. This section provides the details of data creation of the SMILE-College dataset for sentiment analysis.

A. *Student Voice Survey (SVS) Data*

This study uses the SVS response data[2] on the current state of mental health designed by College Pulse, to examine the social and emotional well-being of students and gain insights into their attitudes, preferences, and behaviors. The survey, conducted in 2022, comprised 20 questions and was completed by 2,000 undergraduate students from a panel representing over 1,500 colleges and universities across the United States. Our study focuses on the text responses to the question "***What mental health or wellness services and supports provided by your college are working well? What aspects of mental health and wellness need more attention?***"

[2] https://reports.collegepulse.com/current-state-of-mental-health

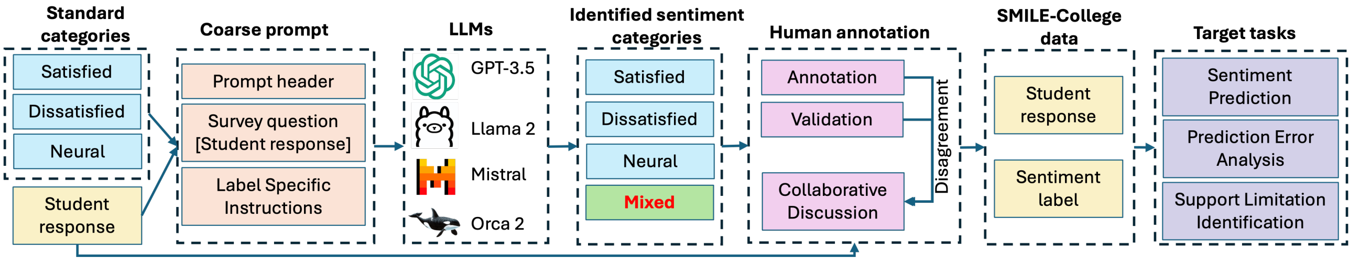


Figure 1: Overall framework for sentiment annotation with human-machine collaboration.

Out of the 2,000 response records, 202 responses represented null values like "n/a", "na", "none", "idk" and empty value. There were also many one word entries that did not provide meaningful answers as shown in first category in Table II. Additionally, several comments only mentioned the name of a service without providing detailed feedback or indicating satisfaction levels (second category of Table II). Also, many responses were overly brief and ambiguous in conveying satisfaction or dissatisfaction as seen in the Irrelevant category, thus impacting the quality of data annotation and model prediction. To ensure the quality of the dataset, we removed this data, by setting a minimum word count of 12 words (based on manual assessment) and removing the irrelevant records during annotation. After this refinement process, we obtained a condensed dataset of 793 records with sufficient context for sentiment analysis.

Table II: Examples of survey records that were removed from the dataset.

Category	Examples
Non-seriousness	• Nsvdejsj • Unknown
Insufficient Information	• eating disorders!!! • tutoring, counseling, and professor care
Irrelevant Information	• while I was home, I felt that school was not worth it as I was home and not doing or going anywhere. • yeah, I'll be there at around noon, and I just got home and I'll get back home from church lol I have a lot of stuff going to my house so I'll

B. *Sentiment Annotation with Human-Machine Collaboration*

Based on the selected data samples, we annotate sentiment labels in a human-machine collaborative manner. While human annotation ensures high accuracy and nuanced understanding, it is costly and time-consuming. Moreover, manually analyzing many samples and defining the appropriate range of sentiment categories becomes challenging. Recently, LLMs offered a scalable solution for annotation [30]. However, there is a significant reduction in performance when transitioning from human labels to LLMs' generated labels due to the inherent noise in the generated labels [31], [32]. Therefore, a viable alternative is to have humans and LLMs work together on this specific annotation task [33].

During our annotation, both LLMs and human annotators contributed unique strengths in a complementary twofold approach. LLMs were first used for quick preliminary analysis, facilitating the identification of sentiment patterns across the entire dataset. By leveraging multiple LLMs, we detected edge cases that suggested the need for an additional sentiment category, enhancing the dataset's granularity. LLMs also helped filter responses with irrelevant or insufficient information, streamlining the annotation process and improving the overall efficiency. Meanwhile, human annotators brought essential depth and contextual understanding of the sentiments, particularly in cases where nuanced interpretation was required. Together, this human-machine collaboration strategy ensured the accuracy and consistency in sentiment annotation and enabled a robust, context-sensitive dataset for analyzing sentiment on mental health support in colleges. The sentiment annotation of SMILE-College can be summarized as the following three steps. Figure 1 provides the overall framework of the annotation.

Step 1. Sentiment Annotation with LLMs. Initially, the number of categories in our data was unclear. To navigate the unstructured nature of the survey responses, we employed Large Language Models (LLMs) to identify response clusters. The goal was to classify the responses into three standard categories: Satisfied (positive class), Dissatisfied (negative class), and Neutral (neutral class). To achieve this, we designed and refined a prompt-engineered approach, leveraging the advanced linguistic capabilities of LLMs. Our strategy involved creating a **coarse prompt** that consisted of three key components:

- Prompt Header: This section contained task-specific instructions, guiding the LLMs to adopt the role of an experienced analyst specializing in mental health text analysis. Here is how we assigned the role in the prompt: "*You are a very experienced analyst trying to analyze the answers to a question asked during a mental health survey. No answer will explicitly mention any of the categories. You have to analyze them based on the rules and categorize them in one word, SATISFIED, DISSATISFIED, or NEUTRAL.*"
- Survey Question and Student Response: This component

Table III: Statistics of the SMILE-College dataset.

	Satisfied	**Dissatisfied**	**Mixed**	**Neutral**
No. of Records	107	376	220	90
Average response length (in words)	21.84	33.01	28.06	18.15
Min response length (in words)	12	12	12	12
Max response length (in words)	93	199	106	46
Average # of sentences in responses	1.89	2.51	2.42	2.03
Min # of sentences in responses	1	1	1	1
Max # of sentences in responses	7	11	8	5

ensured that the LLMs' evaluation was directly informed by the specific content of the survey, grounding its sentiment analysis in the precise context of the student's responses. The survey question is provided in Section III-A.

- Label-Specific Instructions: Comprehensive guidelines were provided for each sentiment category, facilitating accurate categorization of the sentiment expressed in the responses. Guidelines were similar to the criteria for human annotation and validation of sentiment labels in Step 3 of this section.

This zero-shot learning strategy, supported by the coarse prompt, allowed for a preliminary exploration of the spectrum of students' sentiments towards mental health services. By adopting this approach, we were able to harness the capabilities of LLMs to effectively categorize sentiments, despite the initial ambiguity regarding the number of categories.

Step 2. Sentiment Category Identification. We investigated multiple LLMs using the coarse prompt and found a 50.15% prediction agreement for all the models. Among the agreed records, none were classified as neutral, and 70.7% were classified as dissatisfied. Based on our manual review of the predictions and the cases where model outputs diverged, we observed that a significant proportion of the responses contained both positive and negative sentiments, making it difficult to classify them into one of the standard sentiment categories. For example, consider the response "*The student support team at my college has been very helpful in supporting my mental health. I think more attention should be paid on the negative impact of stress from school has on students' mental health.*" Here, the first sentence expresses satisfaction, while the second reflects dissatisfaction. This blend of sentiments highlights the need for a more nuanced annotation approach, as traditional three sentiment categories may not fully capture the complexity of responses in this context.

Inspired by this insight, we shift to a more detailed analysis by introducing "*Mixed*" as a new sentiment category. This transition marks a significant change from a broad to more nuanced sentiment analysis, allowing for a deeper and more precise understanding of the survey responses through human-machine collaboration.

Step 3. Human Annotation and Validation of Sentiment Labels. Based on the four categories of students' sentiments identified in Step 2, we adopt a two-stage human annotation process. The preliminary annotation stage is executed by one annotator, which is subsequently subjected to a validation stage conducted by another annotator. Both annotators are graduate students who are proficient in English. This rigorous process revealed a disagreement rate of 9.98% between the annotations in two phases. These mismatches are resolved collaboratively through collective discussions among the annotators and other researchers involved, leading to a consensus on the final sentiment labels.

The specific criteria for the annotation of each category are as follows. (1) "*Satisfied*": at least 75% of the language expressed satisfaction, with minimal suggestions for improvement. (2) "*Dissatisfied*": at least 75% of the language indicated discontent or suggestions for enhancement, with little mention of satisfaction. (3) "*Mixed*": expressions of satisfaction and dissatisfaction/suggestions were approximately evenly split, with each constituting about 50%. (4) "*Neutral*": no clear emphasis on satisfaction, dissatisfaction, or suggestions for improvement. This discourse focuses on mental health in a college context.

This human-machine collaboration annotation strategy not only enhances this specific sentiment analysis task but also highlights the importance of combining computational analysis with human insights to capture the intricate emotional nuances within students' responses. We enrich the SVS data with the sentiment labels and obtain the SMILE-College dataset for sentiment analysis.

SMILE-College Data Statistics. Table I provides one representative example of each sentiment category. Table III illustrates the basic data statistics of the SMILE-College data. Following data filtering procedures, 266 distinct colleges/universities were covered within the dataset. Notably, the word count distribution across records ranges from a minimum of 12 words to a maximum of 199 words whereas the sentence count ranges from 1 to 11 sentences.

C. *Target Tasks*

To evaluate the usability of the SMILE-College data, we investigated three important tasks, including:

- *Sentiment prediction* (**T1**): text-based multi-class classification of sentiment labels for students' responses with a task-specific fine-grained prompt for LLMs.
- *Prediction error analysis* (**T2**): examine the prediction errors of LLMs across different sentiment categories.
- *Support limitation identification* (**T3**): based on the responses labeled as "Dissatisfied", we utilize the capabilities of LLMs, embedding learning, and clustering techniques to pinpoint the main shortcomings in student mental health support in colleges.

Table IV: Overall sentiment prediction performance and detailed breakdown of predictions across each sentiment category on the SMILE-College *test set*.

	Satisfied			**Dissatisfied**			**Mixed**			**Neutral**			**Overall**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
LR	0.45	0.24	0.31	0.64	0.77	0.70	0.59	0.55	0.57	0.69	0.58	0.63	0.61	0.62	0.61
SVM	0.50	0.33	0.40	0.63	0.78	0.70	0.53	0.40	0.46	0.71	0.63	0.67	0.60	0.61	0.60
BERT	0.68	0.71	0.70	0.88	**0.85**	**0.86**	**0.62**	0.70	0.66	0.88	0.74	0.80	0.79	0.78	0.78
Mistral	0.83	0.48	0.61	0.98	0.52	0.68	0.56	0.50	0.53	0.28	**1.00**	0.43	0.77	0.57	0.60
Orca 2	**0.88**	0.61	0.72	0.95	0.25	0.40	0.34	**0.95**	0.50	**1.00**	0.17	0.29	0.78	0.48	0.46
Llama 2	0.00	0.00	0.00	0.93	0.63	0.75	0.50	**0.95**	0.66	0.52	0.79	0.62	0.65	0.65	0.61
GPT-3.5	0.66	**0.90**	**0.76**	**1.00**	0.75	**0.86**	**0.62**	0.78	**0.69**	0.81	0.89	**0.85**	**0.84**	**0.79**	**0.80**

Table V: Overall sentiment prediction results on *the entire SMILE-College dataset* with zero-shot prompting using LLMs.

	Satisfied			**Dissatisfied**			**Mixed**			**Neutral**			**Overall**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Mistral	0.88	0.28	0.43	0.95	0.46	0.62	0.61	0.64	0.62	0.26	**0.99**	0.41	0.77	0.54	0.57
Orca 2	0.89	0.56	0.69	**0.96**	0.23	0.37	0.32	**0.97**	0.48	**0.93**	0.14	0.24	0.78	0.45	0.42
Llama 2	**1.00**	0.06	0.11	0.89	0.58	0.70	0.51	0.93	0.66	0.54	0.86	0.66	0.76	0.64	0.61
GPT-3.5	0.78	**0.76**	**0.77**	**0.96**	**0.71**	**0.82**	**0.66**	0.90	**0.76**	0.77	0.92	**0.84**	**0.83**	**0.80**	**0.80**

IV. EXPERIMENTS

A. Benchmark Methods

To investigate the three target tasks listed in Section III.C, we perform sentiment analysis by considering **Logistic Regression** (LR) and **Support Vector Machine** (SVM) as baseline models. Subsequently, we also fine-tuned **BERT** for the specific sentiment prediction task. Additionally, we developed task-specific fine-grained prompts for Large Language Models (LLMs) to predict the sentiment labels of student responses. The four LLMs evaluated include: (1) **GPT-3.5** [34], (2) **Mistral** 7 Billion (8-bit quantization) [35], (3) **Llama 2** 7 Billion (8-bit quantization) [36], and (4) **Orca 2** 7 Billion (8-bit quantization) [37]. LLMs enable the classification of the student responses into varying levels of their satisfaction with the mental health services.

The prompt design process is iterative and data-driven to optimize the language models' performance in contextually understanding and analyzing students' survey responses. Following the design and result analysis of the coarse prompt in Section III-B, we develop a fine-grained prompt for sentiment prediction, which consists of the same three components as in the coarse prompt but with four fine-grained sentiment categories (Mixed) and provides specific criteria for each category (see Section III-B).

Since LR and SVM require a training phase, we randomly split the dataset into train, development, and test with the ratios of 0.75/0.05/0.2 and report the results obtained on the test split. The same data split was used for finetuning BERT. To ensure a fair comparison, results for the four LLMs were also obtained from the same test set and provided in Table IV. Additionally, we evaluated the performance of the LLMs on the entire SMILE-College dataset (Table V and Fig. 2).

B. Experimental Setup

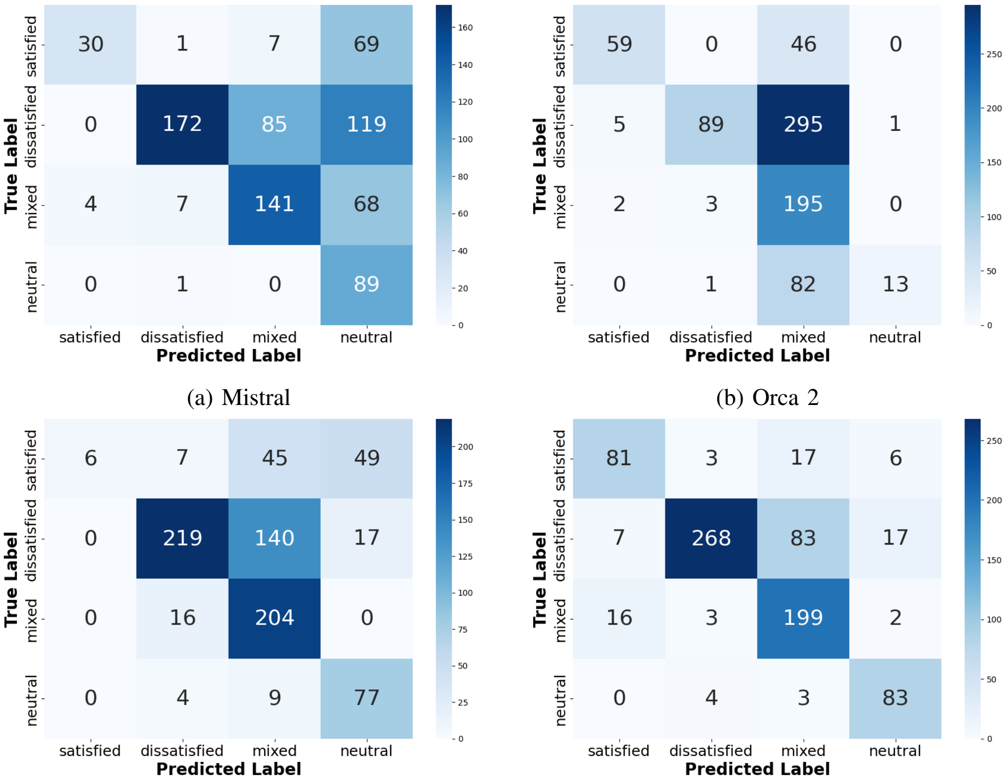
To evaluate the performance of the models on sentiment prediction, We adopt *Precision*, *Recall*, and *F1-score* to evaluate the performance of sentiment predictions. The higher values of these metrics indicate the better performance of a model. We evaluated the overall performance of all sentiment categories with a weighted evaluation to handle the label imbalance issue in the data and ensure a reasonable consideration of all sentiment categories during the evaluation. We use TensorFlow [38] and the Hugging Face library to implement various language models. Our experiments are conducted on a Nvidia Tesla V100 GPU, equipped with 51GB of RAM and 201.2GB of disk space, which has the necessary computational power. During the inference phase, we experiment with 4 or 8-bit for model quantization [39], and with temperature settings ranging from 0.1 to 0.3 to get the best results.

C. Experimental Results

*1) **Sentiment Prediction (T1)**:* Table IV provides the quantitative performance comparison of different models on the test set of the SMILE-College dataset. The best performance of different models is highlighted in bold, while the second-best performance is underlined.

We observe from Table IV that overall GPT-3.5 achieves the highest F1 score of 0.80, outperforming other models. Its large size and robust architecture allow it to deliver a balanced performance across all sentiment categories. The second-best performance is observed from the fine-tuned BERT, with an overall F1 score of 0.78. BERT consistently performs well across most categories, particularly in the Dissatisfied and Neutral categories, where it achieves an F1 score of 0.86 and 0.80, respectively. Its encoder-based architecture continues to be effective in capturing contextual relationships, leading to strong results in sentiment classification.

Interestingly, the other three LLMs, including Mistral, Orca 2, and Llama 2, while more suited for generative tasks, still deliver competitive results when compared to baselines like SVM and LR. This suggests that despite being optimized for text generation, these models exhibit a strong understanding of the contextual intricacies of mental health-related text. For instance, Mistral demonstrates strong recall in the Neutral category (1.00) and Orca 2 exhibits impressive precision in the Dissatisfied (0.95) and Satisfied (0.88) categories. However, Llama 2 underperforms significantly in the Satisfied category, where it fails to produce any meaningful results. This variability suggests that while these models grasp the overall



(a) Mistral

(b) Orca 2

(c) Llama 2

(d) GPT-3.5

Figure 2: Confusion metrics for the four LLMs on *the entire SMILE-College dataset*.

sentiment context, their task-specific performance is not as fine-tuned or consistent as models like BERT or GPT-3.5.

The performance of the four LLM's using zero shot prompting on the entire dataset can be seen in Table V. We observe that GPT-3.5 maintains strong performance with the highest overall F1 score, demonstrating its ability to adapt well in a zero-shot setting without the need for task-specific fine-tuning. Orca 2 demonstrates strong precision but struggles with low recall, particularly in the Dissatisfied and Neutral categories, resulting in a lower overall F1 score. Mistral excels in Neutral recall but suffers from inconsistent precision across categories. Llama 2, notably, performs better for the Satisfied category in the entire dataset (F1 = 0.11) than the test set (F1 = 0.00). Overall, the decoder-based LLMs show consistent performance across both, the test set and the entire dataset.

*2) **Prediction Error Analysis with Confusion Matrix (T2):*** Table V and the confusion matrices in Figure 2 provide a deeper understanding of the LLM's performance variations and error patterns. Across the board, GPT-3.5 shows the most balanced distribution of errors, as reflected by its fewer misclassifications between categories. Notably, there is minimal confusion between the Satisfied and Neutral or Mixed categories, a challenge that other models face more frequently. The matrix reveals that GPT-3.5 handles the overlap between sentiments better than others.

Llama 2 demonstrates the second-best performance in LLMs, excelling in Satisfied sentiment detection with perfect precision. However, frequent misclassifications into Neutral or Mixed categories (Fig. 2(c)), highlight its struggle with recall, leading to an imbalanced performance. Despite this, Llama 2 manages a stronger performance in the Dissatisfied and Mixed categories compared to other models, showing that it can capture negative and mixed emotions fairly well, but lacks the balance needed for more diverse sentiment types.

Mistral and Orca 2 struggle with handling Mixed and Dissatisfied categories. Mistral frequently misclassifies Mixed samples as Neutral or Dissatisfied, while Orca 2 shows confusion between Mixed and Dissatisfied sentiments, though it performs well with Neutral sentiments.

*3) **Support Limitations Identification (T3):*** To enhance the well-being of students, it is crucial to carefully examine the areas of college mental health services that need more attention. We employ GPT-3.5 to identify and extract the limitations based on the survey responses labeled as "Dissatisfied" in the SMILE-College dataset. After manual verification, we obtain the embeddings of the extracted limitations using the sentence transformer [40] and further cluster them using K-Means [41] to systematically categorize the limitations of college mental health services. The examination of each cluster's content reveals predominant themes and topics, which are systematically detailed in Table VI along with the frequencies of the limitations mentioned in all the survey responses. Examining the dataset reveals that the most pressing issue is the quality of counseling services, with 157 mentions,

Table VI: Frequency of Identified Limitations in Mental Health Services by Cluster.

Cluster	Limitations	Freq
1	Quality of Counseling Services	157
2	Availability and Accessibility	76
3	Challenges in accessing the services	76
4	Awareness and Education	73
5	Issues with Therapist Matching	65
6	Inadequacies in support, communication, community connection	64
7	Personal Experiences and Preferences	52
8	Financial and Administrative Concerns	48
9	Diversity and Inclusivity	22
10	Issues with Referrals and Redirection	13

followed by concerns about availability and accessibility, each cited 76 times. These findings highlight the need for colleges to improve counseling services and access. Although less frequent, issues like diversity and inclusivity and referrals also indicate areas for improvement in creating a more inclusive support system.

V. Discussions

Working with real-world student voice survey data presents unique challenges, especially due to the unstructured and often inconsistent nature of student feedback. Data quality is entirely dependent on the respondents' willingness and seriousness to give answers. The open-ended design and subjective nature of survey questions complicated analysis with their broad range of responses. Additionally, inconsistent text generation from decoder-based LLMs made post-processing difficult, limiting the extraction of consistent insights.

Leveraging LLMs offers a significant opportunity to shed light on how mental health support structures are perceived within academic institutions. Additionally, LLMs allow for scalable analysis of subjective data and more personalized mental health interventions, helping shape data-driven policies that better meet student needs and enhance overall mental health services. The ability to highlight recurring issues can prompt institutions to make necessary revisions, improving overall mental health support systems for a more inclusive and effective approach. With this initial exploration of student sentiment on mental health support in colleges using LLMs, we hope to inspire further research into leveraging LLMs to advance mental health-related studies.

In this work, we prioritized ethical considerations, particularly regarding student privacy and potential bias. The Student Voice Survey (SVS) Data, containing students' feedback on mental health services, was already anonymized and de-identified by College Pulse prior to annotation, ensuring privacy protection. To enhance efficiency and accuracy, we employed LLM-based annotations, which were cross-verified by human annotators from different backgrounds. This multilayered approach minimized bias and ensured cultural relevance. Additionally, we transparently documented the role of LLMs in the annotation process and used the publicly available, vetted SVS dataset, aligning with ethical standards for privacy, fairness, and responsible AI use in mental health research.

This work offers significant potential for advancing real-world practices in survey design and data utilization for mental health research. For example, the insights uncovered through the human-machine collaborative annotation process, such as the insufficient or irrelevant survey responses, and the introduction of the "Mixed" sentiment category, underscore the critical role of well-designed survey questions in engaging participants effectively and eliciting more structured, informative responses. Additionally, the SMILE-College dataset's relatively small sample size poses challenges to model performance. Expanding the dataset with additional survey responses in future studies could enhance its robustness and generalizability. Future iterations of the SMILE-College dataset could also incorporate richer annotations, capturing specific issues, benefits, and emotional tone. This includes introducing more granular sentiment categories, such as differentiating dissatisfaction (e.g., service quality vs. accessibility) and satisfaction (e.g., effectiveness vs. convenience), to enable deeper analysis.

VI. Conclusions

Mental health support in colleges and universities is crucial for fostering students' mental health awareness and well-being. However, its effectiveness is hard to evaluate due to various challenges. This paper utilizes student feedback from a public Student Voice Survey, employing advanced LLMs to analyze students' perceptions of college mental health support. A new SMILE-College dataset is created through human-machine collaboration for sentiment analysis. Three important tasks are investigated on the new data, including sentiment prediction, prediction error analysis, and support limitation identification. Experiments reveal that GPT-3.5 performs the best, followed by BERT, in the sentiment prediction task. Additionally, "Quality of Counseling Services" emerged as the most frequently identified limitations faced by students. This data-driven approach facilitates better mental health support evaluation and decision-making.

Acknowledgment

This work was supported in part by the US National Science Foundation grants IIS-2245907, 2047843, and 2437621, and the startup funding from the Stevens Institute of Technology.

References

[1] D. Eisenberg, J. Hunt, and N. Speer, "Mental health in american colleges and universities: variation across student subgroups and across campuses," *The Journal of nervous and mental disease*, vol. 201,1, 2013.

[2] S. K. Lipson, S. Zhou, S. Abelson, J. Heinze, M. Jirsa, J. Morigney, A. Patterson, M. Singh, and D. Eisenberg, "Trends in college student mental health and help-seeking by race/ethnicity: Findings from the national healthy minds study, 2013–2021," *Journal of Affective Disorders*, vol. 306, pp. 138–147, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165032722002774

[3] C. W. Pester, G. Noh, and A. Fu, "On the importance of mental health in STEM," *ACS Polymers Au*, vol. 3, no. 4, pp. 295–306, 2023.

[4] D. J. Drum, C. Brownson, A. D. Burton, and S. E. Smith, "New data on the nature of suicidal crises in college students: Shifting the paradigm," *Professional Psychology: Research and Practice*, vol. 40(3), p. 213–222.

[5] S. K. Lipson, S. Abelson, P. Ceglarek, M. Phillips, and D. Eisenberg, "Investing in student mental health: Opportunities and benefits for college leadership," 2019.

[6] A. Macaskill, "The mental health of university students in the united kingdom," *British Journal of Guidance & Counselling*, vol. 41, no. 4, pp. 426–441, 2013. [Online]. Available: https://doi.org/10.1080/03069885.2012.743110

[7] K. Sonneville, I. Thurston, A. Gordon, T. Richmond, H. Weeks, and S. Lipson, "Weight stigma associated with mental health concerns among college students," *American Journal of Preventive Medicine*, vol. 66, 09 2023.

[8] J. Bautista and S. Schueller, "Understanding the adoption and use of digital mental health applications "apps" among college students: Secondary analysis of a national survey (preprint)," *JMIR Mental Health*, vol. 10, 10 2022.

[9] T. H. M. N. Team, "The healthy minds study - student survey," https://healthymindsnetwork.org/, 2023.

[10] A. C. H. Association, "American college health association national college health assessment," https://www.acha.org/ACHA/Resources/Topics/MentalHealth.aspx, 2024.

[11] M. Priestley, E. Broglia, G. Hughes, and L. Spanner, "Student perspectives on improving mental health support services at university," *Counselling and Psychotherapy Research*, vol. 22, no. 1, 2022.

[12] T. Osborn, S. Li, R. Saunders, and P. Fonagy, "University students' use of mental health services: a systematic review and meta-analysis," *International Journal of Mental Health Systems*, vol. 16, no. 1, p. 57, 2022.

[13] C. Pulse, "College pulse," https://collegepulse.com/, 2024.

[14] T. Adams, M. Moore, and J. Dye, "The relationship between physical activity and mental health in a national sample of college females," *Women & health*, vol. 45, pp. 69–85, 02 2007.

[15] M. Bartlett, H. Taylor, and J. Nelson, "Comparison of mental health characteristics and stress between baccalaureate nursing students and non-nursing students," *The Journal of nursing education*, vol. 55, pp. 87–90, 01 2016.

[16] S. Cleveland, A. Branscum, V. Bovbjerg, and S. Thorburn, "Mental health symptoms among student service members/veterans and civilian college students," *Journal of American college health : J of ACH*, vol. 63, 11 2014.

[17] B. Liu, *Sentiment analysis and opinion mining*. Springer Nature, 2022.

[18] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications," *Knowl. Based Syst.*, vol. 89, pp. 14–46, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:206712020

[19] X. Zhang and Y. LeCun, "Text understanding from scratch," 2016.

[20] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based LSTM for aspect-level sentiment classification," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, J. Su, K. Duh, and X. Carreras, Eds. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 606–615. [Online]. Available: https://aclanthology.org/D16-1058

[21] P. Sood, X. Yang, and P. Wang, "Enhancing depression detection from narrative interviews using language models," in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2023, pp. 3173–3180.

[22] M. Taboada, J. Brooke, M. Tofiloski, K. D. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational Linguistics*, vol. 37, pp. 267–307, 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:3181362

[23] G. Shen, J. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T.-S. Chua, and W. Zhu, "Depression detection via harvesting social media: A multimodal dictionary learning solution," in *International Joint Conference on Artificial Intelligence*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:13959181

[24] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, I. Gurevych and Y. Miyao, Eds. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 328–339. [Online]. Available: https://aclanthology.org/P18-1031

[25] R. Singh and P. Sharma, "An overview of social media and sentiment analysis," in *2021 5th International Conference on Information Systems and Computer Networks (ISCON)*, 2021, pp. 1–4.

[26] K. K. Mohbey, "Sentiment analysis for product rating using a deep learning approach," in *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, 2021, pp. 121–126.

[27] Z. Guo, A. Lai, J. H. Thygesen, J. Farrington, T. Keen, and K. Li, "Large language model for mental health: A systematic review," 2024.

[28] F. Benrouba and R. Boudour, "Emotional sentiment analysis of social media content for mental health safety," *Social Network Analysis and Mining*, vol. 13, no. 1, p. 17, 2023.

[29] A. Shah, R. Shah, P. Desai, and C. Desai, "Mental health monitoring using sentiment analysis," *International Research Journal of Engineering and Technology (IRJET)*, vol. 7, no. 07, pp. 2395–0056, 2020.

[30] Z. Tan, D. Li, S. Wang, A. Beigi, B. Jiang, A. Bhattacharjee, M. Karami, J. Li, L. Cheng, and H. Liu, "Large language models for data annotation: A survey," *arXiv preprint arXiv:2402.13446*, 2024.

[31] J. Mohta, K. Ak, Y. Xu, and M. Shen, "Are large language models good annotators?" in *Proceedings on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models" at NeurIPS 2023 Workshops*, ser. Proceedings of Machine Learning Research, J. Antorán, A. Blaas, K. Buchanan, F. Feng, V. Fortuin, S. Ghalebikesabi, A. Kriegler, I. Mason, D. Rohde, F. J. R. Ruiz, T. Uelwer, Y. Xie, and R. Yang, Eds., vol. 239. PMLR, 16 Dec 2023, pp. 38–48. [Online]. Available: https://proceedings.mlr.press/v239/mohta23a.html

[32] C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang, "Can large language models transform computational social science?" *Computational Linguistics*, vol. 50, no. 1, pp. 237–291, 2024.

[33] S. Wang, Y. Liu, Y. Xu, C. Zhu, and M. Zeng, "Want to reduce labeling cost? GPT-3 can help," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 4195–4205. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.354

[34] J. Ye, X. Chen, N. Xu, C. Zu, Z. Shao, S. Liu, Y. Cui, Z. Zhou, C. Gong, Y. Shen, J. Zhou, S. Chen, T. Gui, Q. Zhang, and X. Huang, "A comprehensive capability analysis of gpt-3 and gpt-3.5 series models," 2023.

[35] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," 2023.

[36] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," 2023.

[37] A. Mitra, L. D. Corro, S. Mahajan, A. Codas, C. Simoes, S. Agarwal, X. Chen, A. Razdaibiedina, E. Jones, K. Aggarwal, H. Palangi, G. Zheng, C. Rosset, H. Khanpour, and A. Awadallah, "Orca 2: Teaching small language models how to reason," 2023.

[38] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[39] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," 2017.

[40] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," 2020.

[41] S. Na, L. Xumin, and G. Yong, "Research on k-means clustering algorithm: An improved k-means clustering algorithm," in *2010 Third International Symposium on Intelligent Information Technology and Security Informatics*, 2010, pp. 63–67.