Emotion Classification Using Natural Language Processing to Support Mental Health Awareness

Aditya Arvind[1], **Robail Yasrab**[2] **& Pramit Saha**[3]

Received March 1, 2026
Accepted July 1, 2026
Electronic access August 15, 2026

Stress, anxiety, and other emotional issues have become more prevalent in modern society. Mental health is a concern surfacing globally, impacting individuals across various demographics. These emotional issues often go unnoticed as people tend to overlook any signals of deteriorating mental health and intense activity overload. This experimental classification study evaluates the technical feasibility of a lightweight fine-tuned LLM for emotional classification using textual semantic analysis. Here, we use AI-driven textual semantic analysis to identify underlying emotions without requiring users to self-report. To train our model, we used the GoEmotions dataset, which contains over 50,000 sentence samples classified by emotion. We trained and fine-tuned the LLaMA model with 1 billion parameters to accurately classify the emotion expressed in a sentence. In addition, we applied 4-bit quantization and Low-Rank Adaptation (LoRA) as fine-tuning techniques to optimize resource and memory bandwidth usage. In our preprocessing pipeline, we analyzed the data and removed duplicates and irrelevant samples from our training process. Ultimately, our training yielded a 62.44% general classification accuracy. These results indicate moderate performance of our model, demonstrating its current limited reliability in emotion classification while also illustrating its potential for improvement with further modifications. Despite thorough development, our research has a few limitations that need further future work. We achieved only 62.44% accuracy, indicating that some element of our approach requires improvement. In addition, the data we used were not specific to any particular group (e.g., adolescents, a major group that faces well-being issues), making this solution rather unspecialized. Although it's generalizable, our research focuses only on textual analysis. In the future, voice analysis could potentially serve as a better approach to inducing awareness of such issues. Also, in our dataset, many classifications were neutral, potentially causing the model to become biased towards this classification. These all serve as experiences that will support improvement in future works. Ultimately, our research showed that the LLaMA model with 1 billion parameters performs moderately well at classifying emotional data but requires modifications to truly improve and be applicable in the real world.

Keywords: Burnout, Mental Health, Emotion

Introduction

Being able to handle our emotions and manage stress helps keep us productive. But the truth is, stress doesn't usually hit all at once; it builds up quietly. In a (study by Sawyer et al.[1]), adolescence is considered the optimal period for developing skills and building a strong foundation for the future. However, the constant stress and overload of daily demands adolescents face inhibit their ability to flourish intellectually, physically, and/or mentally. Furthermore, stressors have a major influence upon mood, behavior, and individuals don't always recognize when they experience frequent overload or why. As exhibited in a (study by McEwen[2]), "it is not just the dramatic stressful events that exact their toll, but rather the many events of daily life that elevate and sustain activities of physiological systems and cause sleep deprivation, overeating, and other health-damaging behaviors, producing the feeling of being stressed out". Individuals often lack the self-awareness to recognize when they are overwhelmed, and the accumulated stress manifests as poor performance (e.g., physical and/or academic/professional) and detrimental emotional regulation. This can eventually lead to serious issues: physical exhaustion and other medical problems. Moreover, finding a solution that addresses these challenges for a variable population with a singular approach is difficult. To that end, concerns of the lack of personalized treatment will arise. That said, advancements in Machine Learning (ML) provide a variety of opportunities for personalized support. For instance, an ML algorithm can use pattern recognition of recent activity (e.g., excessive studying, working and/or insufficient exercise) specific to the user to assess whether the user is

[1] *Northern Highlands Regional High School,*
[2] *University of Cambridge,*
[3] *University of Oxford.*

overwhelmed. Alternatively, implementing textual semantic analysis in a Natural Language Processing (NLP) algorithm can assist in detecting negative emotions in textual responses and may support future emotion recognition systems, pending validation. This intersects with many studies evaluating NLP scoping protocols for screening occupational stress (study by Bieri et al.[3]), broader framework reviews for Depressive disorders (study by Teferra et al.[4]). There are many modern studies focused on deploying passive AI detection strategies for workplace burnout (study by Rana et al.[5]), and systematic research around NLP processing for mental health interventions (study by Malgaroli et al.[6]). In any case, adaptive ML algorithms represent a step forward in this journey.

A pattern recognition algorithm can help assess effort levels during activities (e.g., studying or working long hours). This project aims to share a preliminary technical framework for automated emotion recognition from text. While the model shows potential in classifying emotional patterns, its current performance is not suitable for clinical use. Here, we explore how our modifications and methods enhance the efficiency of our algorithm's development and deployment. Lastly, we aim to detect emotions indirectly: we detect them in written text rather than requiring the user to provide them. This is meant to reduce possible bias in the moment. In addition, the goal is for the algorithms to engage in conversation to help classify their emotions. Specifically, an AI-driven semantic analysis of text identifies emotions without requiring users to self-report their mood, as they may not fully comprehend their emotional state. The purpose of this study is to assess the accuracy of a 1B-parameter (LoRA + 4-bit) model in identifying 28 distinct emotion categories, providing a foundation for future systems. To ensure the model can run on standard consumer hardware, this study utilizes 4-bit Quantization and Low-Rank adaptation (LoRA). Quantization compresses the model's data to reduce memory usage, while LoRA fine tunes only a subset of parameters. These techniques help avoid the need for high cost computing power.

Literature Review

Several prior studies have explored the impact of ML algorithms on emotional and physical health. However, while these studies show potential, there are still several limitations, especially regarding the clinical impact of these models, which may not accurately reflect the results for the issues they address. They serve as learning opportunities that showcase many alternative approaches to project success. A study conducted by Yang et al.[7] collected self-reported symptoms and mood data from students in China. The researchers aimed to determine whether a logistic regression model can predict future symptoms. The results reveal that positive and negative emotion variability, the two main factors, played an essential role in predicting symptoms in a Logistic Regression Model. However, self-reported data may introduce bias due to extreme responses.

A study conducted by Russell et al.[8] aimed to determine whether AI models (e.g., Convolutional Neural Networks) can be paired with acceleration sensors to model and predict physical and cognitive burnout. They analyzed one healthy individual to determine if this methodology works. As a result, this study demonstrated that collaboration between a sensor and a CNN model can provide a remarkably accurate model of physical and mental exhaustion. However, the study's method of testing only one individual raises concerns about the generalizability of the results, as they may not be representative of a larger group. A (study by Penchina et al.[9]) proposed an RNN (Recurrent Neural Network)/LSTM(Long Short-Term Memory) classifier that can recognize states of anxiety in autistic and neurotypical groups using continuous electroencephalography (EEG) signals. The deep learning model had incredible results. For classification between anxious and non-anxious classes, the model achieved an accuracy of 90.82% accuracy. In addition, it also achieved an average accuracy of 93.27%. Although this outcome is exceptional, the study group consisted solely of adolescents. Thus, it will be difficult to generalize these results to adults or to any age group outside those tested. Most of the studies presented above rely on self-reported data, increasing the risk of bias; when subjects provide their own data, responses may be more extreme than what is truly representative of their emotional state. The approach in this study differs: the model, when utilized by a user, does not require the user to state their current mental status. Instead, the model evaluates how users approach particular questions, both general and critical thinking, using those responses to analyze their emotions.

That said, recent advancements in deep learning have transformed the view of automated text-based emotion analysis. Researchers have deployed hybrid learning architectures for generalized psychiatric disorder tracking (study by Aina et al.[10]) and specialized frameworks like nBERT to map subtle emotional trajectories in psychotherapy transcripts (study by Rasool et al.[11]). Newer approaches are being used to evaluate multimodal inputs through multimodal instructional alignments, such as Emotion-LLaMA (study by Cheng et al.[12]) and ensemble multi-transformer voting architectures for depression screening (study by Kasap et al.[13]). Together, these approaches show the shift towards leveraging Large Language Models within mental health scoping initiatives (study by Jin et al.[14]).

Methodology

Our research is an experimental study aimed at evaluating the performance of a fine-tuned model in correctly classifying un-

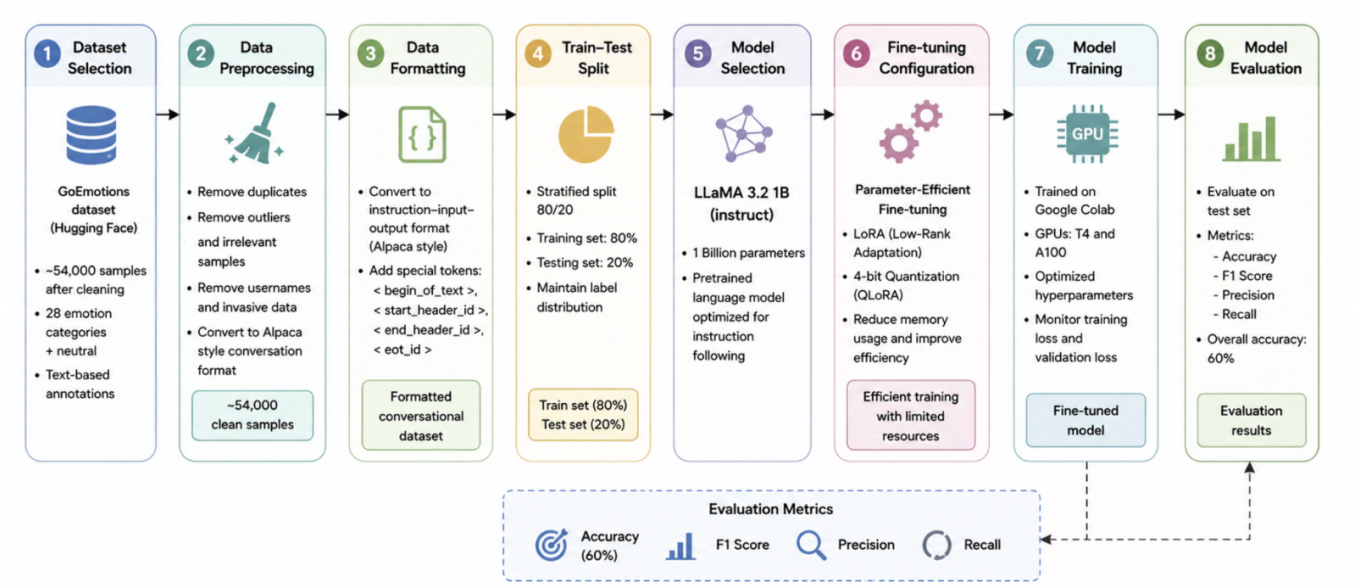

Figure. 1 Visual Abstract - This details the research pipeline: preprocessing the data by removing duplicates/irrelevant samples (GoEmotions dataset), converting the data to model-specific format (Alpaca style), splitting the data (80/20), selecting the model (LLaMA 1B), training the model (LoRA + 4-bit quantization), and evaluating the model (Accuracy, F1, Precision, Recall).

Table 1 Various studies conducted on emotion detection

Study	Dataset Used	Method/Model	Results (Accuracy/F1)	Key Limitation
Penchina et. al (2020)	EEG Signals (Adolescents)	RNN/LSTM	90.82% (Anxiety classification)	Restricted to adolescent autistic/ neurotypical groups
Russell et al. (2021)	Acceleration sensor	CNN	High fatigue prediction accuracy	Tested only on one individual
Yang et al. (2025)	Self -reported mood	Logistic Regression	Predictive of symptoms	Self-reporting bias
Current Study	GoEmotions (Text)	Fine-tuned LLaMa 1B (LoRA + 4-bit)	62.44% (Macro-average accuracy)	Non-specific dataset (not segmented by age group)

derlying emotions in textual data. The goal is to determine how well textual semantic analysis can help with awareness of emotional well-being. We analyzed results after training the model (LLaMA 1B) on textual data from the GoEmotions dataset. This study is intended to establish a baseline for emotion classification performance with a lightweight, fine-tuned model using resource-conscious techniques such as LoRA and 4-bit quantization. We assessed this model mainly based on accuracy, precision, recall, and F1 scores

Data Description

The Kaggle GoEmotions dataset, uploaded to Hugging Face as GoEmotions_Alpaca_Final[15], was used in this research. The dataset's features comprise 28 categories—27 representing emotions and one neutral. The annotation protocol we used in our research was the same as what Demszky et al.[16] used in their GoEmotion dataset. Each of the approximately 54,000 samples had a single classification from the 28 presented categories. Given a text, whether a few words or a sentence, the task was to predict the emotion it conveys. Despite the feasibility of this approach, several ethical considerations were taken into account. Initially, voice recognition was preferred over textual analysis. However, collecting adequate voice data for this project would have raised significant ethical concerns, such as potential privacy invasion if the model inadvertently leaked personal voice data. To reduce such concerns and mitigate ethical risks, a model based on textual rather than voice analysis was developed.

Preprocessing

Although the initial dataset had around 58,000 samples, the outliers and excess irrelevant samples were removed. We dis-

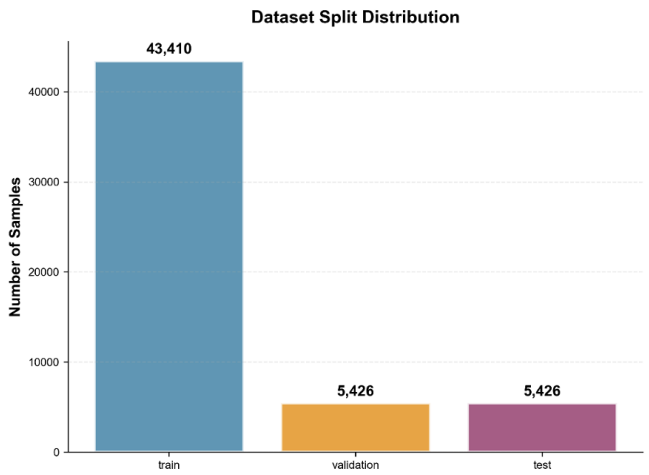

Figure. 2 This bar chart shows the raw number of samples in the dataset for each part of the process: training, validation, and testing.

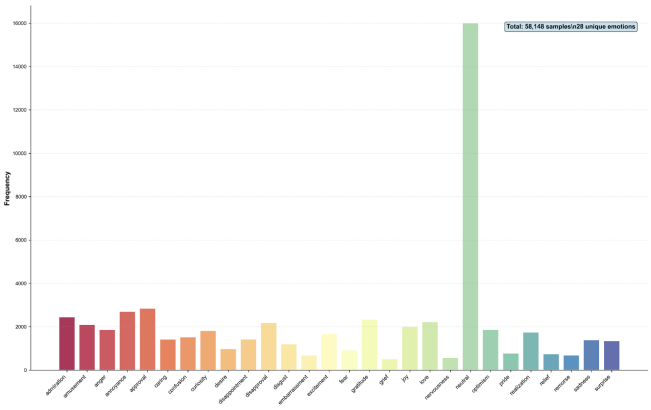

Figure. 3 This chart shows the frequency of each of the 28 emotions in the dataset

carded about 4,000 (7%) multi-label entries and used only single-label entries. Because this study was designed as a single-label classification task, comments with more than one emotion label were excluded rather than collapsed into one label. No tie- resolution rule was applied because multi-labels were not converted; they were removed. This resulted in a polished dataset of about 54,000 (93%) samples (post-filter distribution shown in Figure 3). Despite these changes, the features in the original dataset remained unchanged. The original 28 categories — 27 emotions and 1 neutral — were also applied to the cleaned dataset. To adapt this dataset for an LlaMA model (e.g., LlaMA-3.2 1B), we reformatted it into dialog format. The structured conversation format uses special tokens that the model is trained on. It uses tokens like ‘< begin_of_text >, < start_header_id >, < end_header_id >, and <eot_id > to structure turns in a conversation between user, assistant, and system. For our training processes, we leveraged an 80/20 split for training and testing data. We proposed this split because the model can gain exposure to a wide variety of examples for training — 80% of the samples—while still receiving sufficient assessment—20% of the data is allocated to testing the model. In terms of the dataset, the GoEmotions dataset we used was already partially cleaned (e.g., removal of the majority of vulgar words and general duplication removal), and our conversion of the dataset to Alpaca style further formatted it: we removed usernames and other unnecessarily invasive data that were not required for the goal/task

Model Selection

The unsloth/Llama-3.2-1B-Instruct-bnb-4bit model[17] was selected for fine-tuning. This approach was found to be highly suitable, as it effectively fulfills the intended purpose while requiring minimal resources. Compared to other Llama models with 7 or 13 billion parameters, this model trains much faster and is significantly more efficient at emotion classification. A model for emotion classification does not necessarily need widespread general knowledge—a feature relevant to larger models. Instead, an efficient model for analyzing language and emotion patterns is preferable for this task.

The model initially used 32-bit weights, which proved inefficient and superfluous for this application. Deploying optimization routines for granular datasets often presents computational issues. To work around this, parameter-efficient methods such as low-rank matrices or low-bit quantization are leveraged to map subtle emotional trajectories across downstream behavioral health datasets (study by Kermani et al.[18]). Therefore, 4-bit quantization was applied, reducing the model's parameter size to 4 bits. The 4-bit quantized model was selected to optimize memory bandwidth and fine-tune the model. Additionally, the model was already trained on instructional data, demonstrating its utility for emotion classification and for recognizing conversation patterns. This strategy is similar to implementations evaluating parameter efficient LLaMA and LoRA fine-tuning combinations for custom emotional mapping (study by Yacoubi et al.[19]). By using these methods and framework, developers can construct an AI powered mental health utility that prioritizes local execution for data privacy preservation (study by Bagane et al.[20]). This builds upon fundamental methodologies using deep learning configurations for text-based emotion indexing (study by Bharti et al.[21]) and extraction of affective vocabulary signatures to detect depressive burnout variants (study by Haug et al.[22]).

The model's ultimate capabilities are what make it valuable for our purpose. It is faster and uses fewer resources for text analysis, demonstrating remarkable token efficiency. Also, it

is proficient at identifying emotions in text, which is directly related to our final goal. Finally, it is strikingly compatible with LoRA and Unsloth fine-tuning. Both of these features enable faster training and more efficient use of resources while maintaining outcome quality

Implementation

The project implementation was conducted entirely in Python, a programming language commonly used for artificial intelligence applications. Code was developed and tested in Google Colab, providing a user-friendly environment for tracking resource usage and ensuring the effectiveness of resource-saving processes. Several Python libraries were utilized to enhance model performance, including SciKit-Learn for evaluation metrics, PyTorch for main code execution, and Matplotlib for visualizing various project features.

Google Colab's A100 GPU was used for the main code execution, while the T4 GPU was used for testing smaller components of the processes. The CPU provided by Colab was avoided, as it proved ineffective for training and testing on models of this size (1B parameters). To further ensure resource efficiency, LoRA adapters were trained and implemented. This approach freezes the base model's parameters and incorporates new adapters for task-specific tuning—in this case, emotion classification. Finally, Ollama was used to host and run the model locally, providing a user-friendly interface similar to commonly used LLMs.

Results

Performance Metrics

Accuracy, precision, recall, and F1 score are crucial metrics for evaluating the trained model. We macro-averaged all the metrics across all 28 emotion classes.

Accuracy score: Describes how often the model's predictions were correct out of all predictions

Macro Precision Score: Describes the average proportion of correct classifications per emotion class

Macro Recall Score: Describes how often the model correctly classified an emotion out of all occurrences of that class in the dataset

Macro F1-Score: Balance of both the Precision and Recall scores

Results (Metrics)

A right-tailed binomial probability significance test yielded statistically significant results. For context, the probability that the model classifies correctly through random guesses is 1/28 = 3.57%. The significance test was conducted to determine the probability that the model achieved an accuracy of 62.44% if it were truly guessing randomly. Since the probability of this happening was below any reasonable significance level, we can reject the hypothesis that the model achieved this accuracy by simple guessing, supporting the conclusion that the model learned patterns in the data.

Table 2 Final Metrics Wilson Score method, $n = 4,545$

Metric	Score	95% CI
Accuracy	0.6244	(0.6102, 0.6384)
Macro Precision	0.5395	(0.5250, 0.5540)
Macro Recall	0.5080	(0.4935, 0.5225)
Macro F1	0.4964	(0.4819, 0.5109)

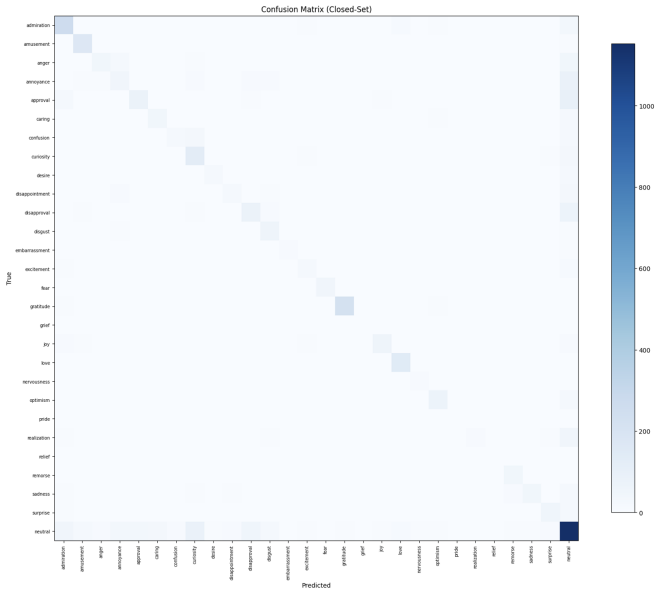

Figure. 4 This is the confusion matrix for the 28-emotion classification task.

This diagram shows how often the model correctly predicted an emotion for each class. We observe a strong off-diagonal pattern in this confusion matrix. One unique pattern we see is that the model classified neutral correctly most of the time, while performing poorly on other classes.

Ultimately, we observe that the model performed moderately, with 62.44% accuracy and a macro F1 score of 0.4964. The resulting macro F1 score demonstrates that the model performed inconsistently across all classes. The accuracy indicates that the model sometimes correctly classified emotions but often made incorrect predictions. As we saw in our confusion matrix, there was a pronounced off-diagonal pattern, further supporting the idea of poor performance across most classes. However, one key issue we identified is that the model

Figure. 5 The diagram above shows the F1 scores per class for emotion classification, with 95% confidence intervals calculated using the Wilson score method. As shown in the diagram, the model's performance varies significantly across classes, and the dotted red line indicates the macro F1 score.

performs significantly better when predicting the neutral class than when predicting other classes.

As previously mentioned, the F1 score indicated substantial inconsistency in the model's performance across all classes. We see decent performance in classes such as gratitude, amusement, admiration, and love. This indicates that the model often predicted these correctly. On the contrary, we observe many classes in which the model performed relatively poorly. These classifications include grief, relief, pride, realization, and annoyance, which had particularly low F1 scores. Overall, this model shows some potential, as it was able to classify certain emotions correctly. However, there is a clear need for improvement, calling for modifications that can enhance the model's overall performance across all metrics and potentially be applicable in the real world.

This is where our algorithm is integrated with Ollama and runs locally on our computer.

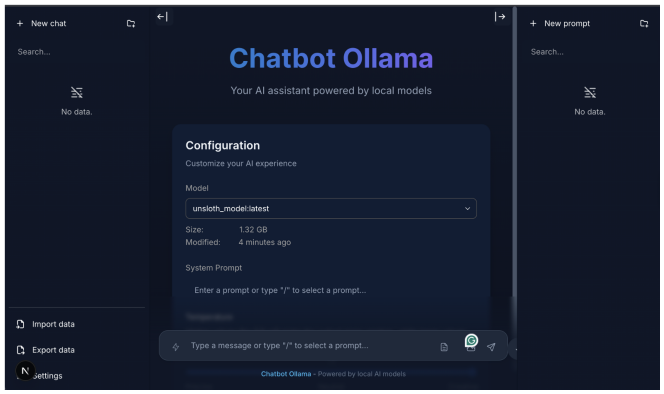

Figure. 6 This image is the final product, depicting the complete user interface.

Discussion

The overall results show decent performance, with 62.44% accuracy and a macro F1 score of 0.4964. This indicates that the model correctly classified the emotions more than half the time. However, for over one-third of the classifications, the model's classification was incorrect and did not match the true emotion presented in the text. The model's inconsistency across all classes is also an issue. This suggests that the model was more accustomed to some emotions than others. This led to many classification errors across different labels, as shown in the confusion matrix. The model's over-specialization to certain classifications, particularly neutral, is most likely due to their higher prevalence in the dataset. As shown in Figure 3, neutral constituted the majority of the data, and other classifications also partially overshadowed minority classes. Overall, the dataset's overrepresentation of certain emotions induces this bias in the model. In addition, the model may perform better on some emotion classifications because they are presented more clearly in the given text, whereas other emotions are more vague, making them harder to recognize. Furthermore, because the GoEmotions dataset consists of gen-

eral Reddit comments, these results cannot be generalized to a specific group. Future work must include a dataset that can be segmented by age, emotion type, and other relevant factors to validate the model's effectiveness.

Limitations

Despite thorough process development and research, our approach still had multiple limitations and issues. Adolescents are among the largest groups facing the same well-being issues we are trying to address and raise awareness of. However, the dataset we used to train our model was not adolescent-specific. The comments were from general Reddit posts, not separated by age group. This means we do not have sufficient evidence to prove that this works particularly well for a specific group. Another major issue we faced is that the accuracy and macro F1 score were only 62.44% and 0.4964, respectively. These low performances demonstrate that this potential solution is not yet ready for real-world applications due to its limited reliability. One reason for this limited reliability is that we restricted this model's scope to interpreting textual data. If we train a model using similar methods but focus on voice analysis, we may obtain more accurate data, thereby yielding a more reliable product. We also did not perform cross-validation, which would further substantiate the model's status on reliability. This study did not include a full fine-tuning baseline or a direct GPU memory comparison, so the use of LoRA and Quantization is limited to a resource-conscious option rather than experimentally validated efficiency improvements. Ultimately, these limitations serve as lessons for future work to help us focus our solution on adolescents, establish an equally representative dataset, and more rigorously evaluate the model to ensure its robustness in real-world applications.

Conclusion

This project demonstrated a technical framework for automated emotion detection. While the model shows potential in classifying general text, its current accuracy is not suitable for medical purposes or burnout prevention in specific groups like adolescents etc. A technology-focused approach was adopted to develop an AI-driven algorithm capable of detecting specific emotions. This provides individuals with an option to identify their emotions, preventing perception bias—bias introduced by information received through a warped perspective—while also reducing the cognitive effort required to obtain an emotion classification. This is a baseline user-friendly tool that may support emotion recognition systems pending further validation.

The dataset consisted of approximately 54,000 samples, an optimal quantity for processing with a 1B-parameter model. However, this led the model to be trained and evaluated more frequently on certain words, since some emotions occurred more often in the dataset than others. Despite this minor inconsistency, no significant issues were observed, and all quantities—dataset size, model parameter size, and train/test split (80/20)—were ideal for the emotion/text classification task.

Future work should focus on identifying alternative datasets with significantly larger sample sizes, thereby substantially improving model accuracy as the volume of training and testing data increases. Additionally, using GPUs more efficient than Google Colab's T4 and A100—such as personal GPUs—would enable more efficient model training and eliminate issues related to usage limits. Resource-saving procedures have already been deployed; further adjustments to hyperparameters (e.g., batch size and epoch size) will be crucial for optimizing training speed and accuracy. Expanding the application scope is also a priority. For example, integrating physical attachments with the model could enable more efficient tracking of adolescents' emotions. Several sources have inspired this direction, notably a (study by Nghiem et al.[23]) from Cornell's People-Aware Computing Lab, which used physical sensors for passive data retrieval. Such approaches motivate further expansion of impact in this area of research.

Code Availability

The code for this project is available at Link

Acknowledgements

Guidance and mentoring by PhD. Researchers Robail Yasrab and Pramit Saha, affiliated with CCIR from University of Cambridge and Oxford University

References

1 S. M. Sawyer and R. A. Afifi and L. H. Bearinger and S.-J. Blakemore and B. Dick and A. C. Ezeh and G. C. Patton, Adolescence: A Foundation for Future Health. The Lancet. Vol. 379, No. 9826, pp. 1630–1640, 2012., https://doi.org/10.1016/S0140-6736(12)60072-5.

2 B. S. McEwen, Protective and Damaging Effects of Stress Mediators: Central Role of the Brain. Dialogues in Neuroscience. Vol. 8, No. 4, pp. 367–381, 2006., https://doi.org/10.31887/DCNS.2006.8.4/bmcewen.

3 J. S. Bieri and C. Ikae and S. B. Souissi and T. J. Müller and M. C. Schlunegger and C. Golz, Natural Language Processing for Work-Related Stress Detection Among Health Professionals: Protocol for a Scoping Review. JMIR Research Protocols. Vol. 13, p. e56267, 2024., https://doi.org/10.2196/56267.

4 B. G. Teferra and A. Rueda and H. Pang and R. Valenzano and R. Samavi and S. Krishnan and V. Bhat, Screening for Depression Using Natural Language Processing: Literature Review. Interactive Journal of Medical Research. Vol. 13, p. e55067, 2024., https://doi.org/10.2196/55067.

5 R. Rana and N. Higgins and T. Stedman and S. March and D. F. Gucciardi and P. D. Barua and R. Joshi, Passive AI Detection of Stress and Burnout Among Frontline Workers. Nursing Reports. Vol. 15, No. 11, p. 373, 2025., https://doi.org/10.3390/nursrep15110373.

6 M. Malgaroli and T. D. Hull and J. M. Zech and T. Althoff, Natural Language Processing for Mental Health Interventions: A Systematic Review and Research Framework. Translational Psychiatry. Vol. 13, No. 1, p. 309, 2023., https://doi.org/10.1038/s41398-023-02592-2.

7 Y. Yang and J. Wang and H. Lin and X. Chen and Y. Chen and J. Kuang and Y. Yao and T. Wang and C. Fu, Emotion Dynamics Prospectively Predict Depressive Symptoms in Adolescents: Findings from Intensive Longitudinal Data. BMC Psychology. Vol. 13, Article 386, 2025., https://doi.org/10.1186/s40359-025-02699-9.

8 B. Russell and A. McDaid and W. Toscano and P. Hume, Predicting Fatigue in Long Duration Mountain Events with a Single Sensor and Deep Learning Model. Sensors. Vol. 21, No. 16, p. 5442, 2021., https://doi.org/10.3390/s21165442.

9 B. Penchina and A. Sundaresan and S. Cheong and A. Martel, Deep LSTM Recurrent Neural Network for Anxiety Classification from EEG in Adolescents with Autism. In: M. Mahmud, S. Vassanelli, M. S. Kaiser and N. Zhong (Eds.), Brain Informatics (BI 2020). Lecture Notes in Computer Science. Vol. 12241, pp. 227–238. Springer, 2020., https://doi.org/10.1007/978-3-030-59277-6_21.

10 J. Aina and O. Akinniyi and M. M. Rahman and V. Odero-Marah and F. Khalifa, A Hybrid Learning-Architecture for Mental Disorder Detection Using Emotion Recognition. IEEE Access. Vol. 12, pp. 91410–91425, 2024., https://doi.org/10.1109/ACCESS.2024.3421376.

11 A. Rasool and S. Aslam and N. Hussain and S. Imtiaz and W. Riaz, nBERT: Harnessing NLP for Emotion Recognition in Psychotherapy to Transform Mental Health Care. Information. Vol. 16, No. 4, p. 301, 2025., https://doi.org/10.3390/info16040301.

12 Z. Cheng and Z.-Q. Cheng and J.-Y. He and K. Wang and Y. Lin and Z. Lian and X. Peng and A. Hauptmann, Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Instruction Tuning. Advances in Neural Information Processing Systems (NeurIPS). Vol. 37, 2024., https://doi.org/10.52202/079017-3518.

13 F. Kasap and S. Omurca and E. Ekin, Ensemble of Transformers for Depression Emotion Classification. Cognitive Neurodynamics. 2026., https://doi.org/10.1007/s11571-026-10444-0.

14 Y. Jin and J. Liu and P. Li and B. Wang and Y. Yan and H. Zhang and C. Ni and J. Wang and Y. Li and Y. Bu and Y. Wang, The Applications of Large Language Models in Mental Health: Scoping Review. Journal of Medical Internet Research. Vol. 27, p. e69284, 2025., https://doi.org/10.2196/69284.

15 AA65327, GoEmotions_Alpaca_Final [Data Set]. Hugging Face. Retrieved December 22, 2025., https://huggingface.co/datasets/AA65327/GoEmotions_Alpaca_Final.

16 D. Demszky and D. Movshovitz-Attias and J. Ko and A. Cowen and G. Nemade and S. Ravi, GoEmotions: A Dataset of Fine-Grained Emotions. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4040–4054, 2020., https://doi.org/10.18653/v1/2020.acl-main.372.

17 Unsloth, Llama-3.2-1B-Instruct-bnb-4bit [Model]. Hugging Face. Retrieved December 22, 2025., https://huggingface.co/unsloth/Llama-3.2-1B-Instruct-bnb-4bit.

18 A. Kermani and V. Perez-Rosas and V. Metsis, A Systematic Evaluation of LLM Strategies for Mental Health Text Analysis: Fine-Tuning vs. Prompt Engineering vs. RAG. In Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology, pp. 172–180, 2025., https://doi.org/10.18653/v1/2025.clpsych-1.14.

19 I. Yacoubi and R. Ferjaoui and W. E. Djeddi and A. B. Khalifa, Advancing Emotion Recognition through LLaMA3 and LoRA Fine-Tuning. 2025 IEEE 22nd International Multi-Conference on Systems, Signals & Devices, pp. 348–353, 2025., https://doi.org/10.1109/SSD64182.2025.10989922.

20 P. Bagane and A. Dahiya and S. Kacheria and A. Sehgal and S. Bajpai and O. A. Jebessa, AI-Powered Mental Health Application with Data Privacy Preservation. MethodsX. Vol. 16, p. 103756, 2025., https://doi.org/10.1016/j.mex.2025.103756.

21 S. K. Bharti and S. Varadhaganapathy and R. K. Gupta and P. K. Shukla and M. Bouye and S. K. Hingaa and A. Mahmoud, Text-Based Emotion Recognition Using Deep Learning Approach. Computational Intelligence and Neuroscience. p. 2645381, 2022., https://doi.org/10.1155/2022/2645381.

22 S. Haug and M. Kurpicz-Briki, Burnout and Depression Detection Using Affective Word List Ratings. Studies in Health Technology and Informatics. Vol. 292, pp. 43–48, 2022., https://doi.org/10.3233/SHTI220318.

23 J. Nghiem and D. A. Adler and D. Estrin and C. Livesey and T. Choudhury, Understanding Mental Health Clinicians' Perceptions and Concerns Regarding Using Passive Patient-Generated Health Data for Clinical Decision-Making: Qualitative Semistructured Interview Study. JMIR Formative Research. Vol. 7, p. e47380, 2023., https://doi.org/10.2196/47380.