Multi-Label Classification of Suicidal Thoughts

Hyunjin Lim[1] and Gubeom Kwon[#]

[1]Daegu International School, Republic of Korea
[#]Advisor

ABSTRACT

Teenage suicide rates continue to rise and given teenagers' increased communication via digital platforms, traditional identification techniques can overlook subtle indicators of distress. This study uses artificial intelligence to examine social media text data for early identification of teenage suicidal thoughts. Using multi-label classification, we optimized the BERT (Bidirectional Encoder Representations from Transformers) model to handle complex cases where people may display several risk variables concurrently. Filtered for relevance to student life, the dataset consisted in student-generated content from Reddit's SuicideWatch and Humor subreddits, tagged across academic performance, bereavement, psychological problems, family issues, and past suicide attempts. Particularly in mental illnesses (F1 = 0.95) and family issues (F1 = 0.91), the model demonstrated great detection ability, high precision, recall, and F1-scores. This method presents a rich analysis that could improve early intervention and support for teenage mental health problems by emphasizing the possibilities of integrating machine learning into real-time monitoring systems.

Introduction

According to National Alliance on Mental Illness, nearly one fifth of American teenagers report serious thoughts of suicide, and half of these have made an attempt to do so. Between 1999 and 2020, over 47,000 American adolescents lost their lives to suicide, with sharp increases year by year, especially among girls and minority youth. Teenagers are increasingly vulnerable, as their prefrontal cortex is being developed, which impacts judgment and decision-making; hence, they are more prone to make impulsive decisions, including those resulting in self-harm or suicide (Ravindranath et al., 2024).

Traditional methods of identifying students who are at risk, such as surveys and counseling, usually rely on self-reporting and face-to-face interactions, which may not fully capture the varied and sometimes hidden signs of distress, especially in a generation that communicates primarily through digital platforms. Thus, a possible solution is to integrate machine learning models that can analyze large volumes of text data from mainly social media sites such as the Reddit dataset; such models allow quicker and targeted treatments that can perhaps save lives by detecting subtle indicators of depression and suicidal ideations (Arowosegbe and Oyelade, 2023).

Artificial intelligence (AI) would be a solution for such challenges by allowing analysis of large volumes of text data from various sources, which includes social media; for example, Deshpande and Rao demonstrated how sentiment analysis uses social media data and how the Multinomial Naive Bayes classifier manages text input better than Support Vector Machines (SVM) (Deshpande and Rao, 2018). Similarly, Babu and Kanagaused emotion AI to analyze Twitter feeds by employing natural language processing (NLP) and machine learning techniques, such as SVM and Naive Bayes classifiers (Babu and Kanaga, 2022). As a result, emotion AI can classify texts into emotional states that reflect underlying depression or suicidal ideation.

Beyond basic classifiers, Aldhyani et al. examined text from Reddit's SuicideWatch using more advanced models such as CNN-BiLSTM and XGBoost; similarly, Allayla and Ayvaz took this a step further by proposing a large-scale data architecture for real-time suicide detection using a Multilayer Perceptron (MLP) classifier (Aldhyani et al., 2022; Allayla and Ayvaz, 2024).

BERT (Bidirectional Encoder Representations from Transformers) is a language modell which revolutionized text analysis due to its capacity to capture intricate language patterns; initially, Van Der Lee emphasized how BERT can handle complex human-written language patterns, while Diniz et al. created Boamente, a BERT-based approach to track suicidal ideas through smartphone text data (Van Der Lee, 2022; Diniz et al., 2022).

We used the BERT model, specifically the "bert-base-uncased" checkpoint, for its contextualized rich language understanding capability and employed fine-tuning to adapt the model to our specific task of identifying suicidal ideation and related mental health indicators. We hypothesize that by fine-tuning a language model on student-generated text data, we can develop a model capable of accurately identifying signs of suicidal ideation.

Additionally, we specifically used multi-label classification, where an instance can belong in multiple categories at the same time. We also hypothesize that by employing multi-label classification, our model will not only detect suicidal ideation but also identify potential contributing factors, such as academic stress, family issues, or mental disorders. Therefore, a major progress in early detection and care for mental health problems among teenagers could come from including machine learning methods into real-time monitoring systems (Thakkar et al., 2024).

Materials and Methods

This model underwent training in Python within a Jupyter notebook setup; specifically, we used PyTorch as the backend and the HuggingFace "Transformers" package for multi-label categorization; we selected such configurations due to its strong support for deep learning operations and seamless integration with the HuggingFace library, in which we can effectively use modern machine learning tools and frameworks.

An integral part that we have done is fine-tuning, in which we modify a pre-trained model (in this case, BERT) to more accurately identify indications of stress in student-generated content; this fine-tuning technique allows the model understand the intricate details of the data, especially small variations in language that may suggest mental health problems (Devlin et al., 2019; Rogers et al., 2020). We also improved the BERT model's accuracy to correctly identify at-risk people within student life by adjusting it on a carefully selected and filtered dataset.

Using the "bert-base-uncased" checkpoint enhances its capacity to read and analyze complicated text input, and since the model is trained on a large corpus of English text, the model begins with a strong pre-trained language knowledge. The learning rate (set to 2e-6), an important hyperparameter that controls the step size during gradient descent, is kept small to enable precise fine-tuning without overshooting optimal parameters. Both the training and evaluation batch sizes, which are set to 8, determine the number of samples processed before the model is updated, with smaller batch sizes allowing for more precise model updates, desipte the cost of longer training time; also, to further stabilize training, we used 100 warmup steps to gradually increase the learning rate. Additionally, we applied a weight decay of 0.01 in order to regularize the model and limit the risk of overfitting by not allowing excessively large weight values.

Additionally, after tokenization, we used a data collator to adjust the padding of the tokenized sequences, which means that all sequences within a batch are of uniform length by adding padding tokens where necessary, as it allows the model to handle multiple sequences simultaneously without errors due to differing sequence lengths. The loss function was cross-entropy loss, while the AdamW optimizer was used for parameter optimization.

We first scraped data from the Suicide Watch Subreddit and Humor Subreddit, covering posts from 2009 to 2020, and collected a total of 232,074 entries from various sources in Kaggle. The Humor Subreddit data was used to obtain non-suicidal data. After scraping, we applied keyword filters focusing on student-related terms such as "school," "exam," "project," "test," "AP," "IB," "teacher," and "parent" to reduce the dataset; this filtering step narrowed the dataset to approximately 20,000 entries. Then, we manually filter the data to determine if the content was written by a student, refining it to 10,184 entries for model training. Specifically, three sections made up the dataset: 1018 entries for validation (examined every 100 steps during training), 8147 entries for training, and 1019 clean entries for testing that was not utilized in training. This split allowed a strong evaluation of the model through the use of data it had not encountered during training, which would guarantee its performance. Data labeling covers academic performance, death of a loved one, psychological disorders including anxiety, family-related issues, and past suicide attempts; every category was given a corresponding label, and entries were allowed to have several labels, as for some of the texts, it is possible to be in more than one category (Orozco et al., 2018; Grimmond et al., 2019; Bilsen, 2018). Originally assigned to distinct columns, labels were later combined into a single column to match Hugging Face's Transformers library's training arrangement; additionally, labels for entries that conveyed a general sense of distress without specific reasons, such as "no specific reason but clearly distressed," were also included.

Text submissions were first tokenized with the BERT tokenizer, which transforms them into a format suitable for BERT model input. Texts longer than this limit would be truncated, which would lead to potential loss of critical information, so they were excluded to ensure the model could accurately understand the entire content. We excluded any text entries exceeding 5000 characters, as the BERT model has a maximum token limit of 512 tokens, which approximately corresponds to 4,500 characters.

The sigmoid function is used to convert the model's raw outputs (logits) into probabilities, which are then used to make binary predictions for each label. The resulting binary predictions are then used to compute final classification metrics, such as accuracy, F1 score, precision, and recall.

Results

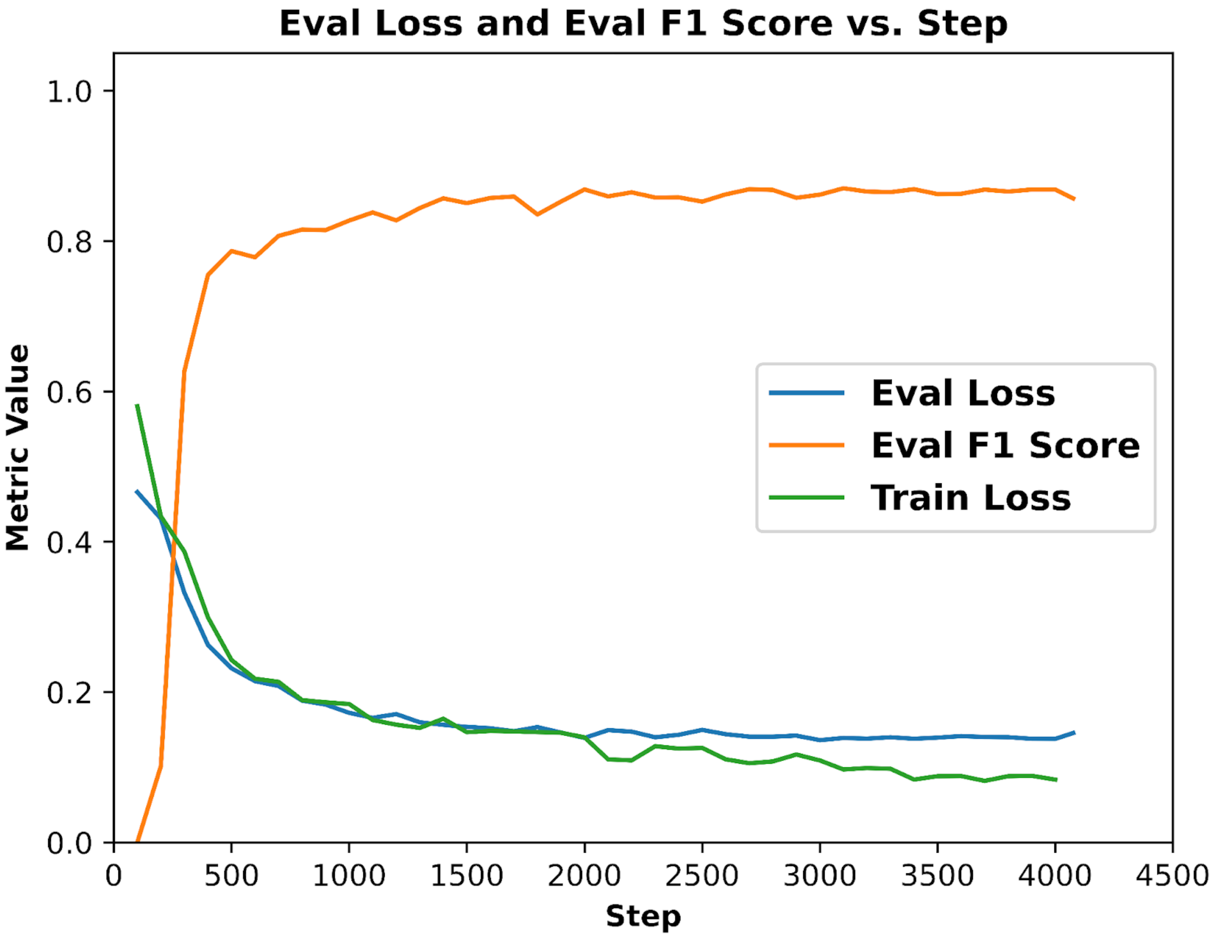

Figure 1. Training and evaluation performance metrics across 4076 steps. This figure illustrates the performance of the model over 4076 training steps, measured at intervals of 100 steps per batch of 8 during a 4-epoch training process. The graph shows the evaluation loss (blue line), training loss (green line), and evaluation F1 score (orange line).

We used multi-label classified text data to train and evaluate the base model by fine-tuning, and, in order to do so, we analyzed the accuracy, F1-score, precision, and recall to test our model's performance to make sure the model is assessed accurately in every category. Reflecting the model's capacity to prevent false positives, precision calculates the proportion of actual positive predictions among all positive predictions. Recall indicates the ability of the model to catch all relevant instances by evaluating the proportion of true positives among all actual positives. The F1-score is the harmonic mean of recall and accuracy, which accounts for a fair evaluation of the model's performance. Accuracy measures the proportion of correct predictions among all predictions.

Among these measures, we concentrated mostly on the F1-score since it provides a good balance between precision and recall and is appropriate for our multi-label classification work where both false positives and false negatives are important. Although the F1 score stabilizes after a small amount steps, the model's losses steadily decrease, which indicates that learning was successful – the constant decline in both training and evaluation loss without significant deviation between them suggest that the model is neither underfitting nor overfitting (Figure 1). While overfitting would result due to an apparent distinction between training and evaluation loss, suggesting the model's inability to generalize data, underfitting would result due to the model failing to capture the patterns in the training data, which leads to high training loss; also, the stability displayed in the graph suggests that our final model has not overfitted to the training set but rather has generalized well to the unseen data.

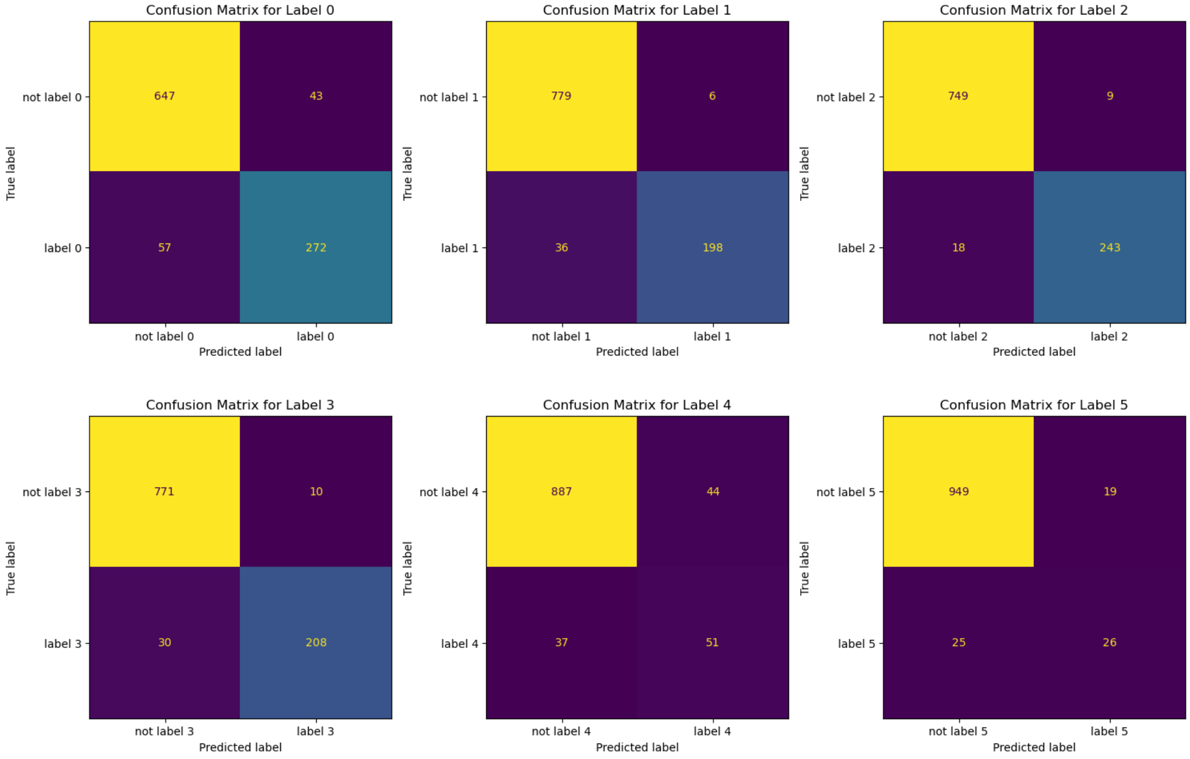

Figure 2. Confusion matrices for multi-label classification of suicidal indicators across various categories. This figure presents the confusion matrices for six different labels used in the multi-label classification of mental health issues for test data, which consists of 1019 rows: academic performance (Label 0), death of a loved one (Label 1), psychological

disorders including anxiety (Label 2), family-related issues (Label 3), past suicide attempts (Label 4), and other suicide attempts (Label 5). At least one thousand different texts were used to prevent inaccuracies in multi-label classification.

Table 1. Precision, Recall, and F1-Score for Multi-Label Classification of Suicidal Indicators Across Various Categories. This table presents the performance metrics for six different test data types in the multi-label classification model: academic performance, death of a loved one, psychological disorders including anxiety, family-related issues, past suicide attempts ("second"), and other suicide attempts. The table lists precision, recall, F1-score, and support (number of instances) for each category. High precision and recall scores are observed for categories such as psychological disorders (Mental, F1 = 0.95) and death of a loved one (Death, F1 = 0.90).

	Precision	**Recall**	**F1-Score**	**Support**
Academic	0.86	0.83	0.84	329
Death	0.97	0.85	0.90	234
Mental	0.96	0.93	0.95	261
Family	0.95	0.87	0.91	238
Second	0.54	0.58	0.56	88
Other	0.58	0.51	0.54	51
Weighted Avg	0.89	0.83	0.86	1201

We created a confusion matrix for each of the labels (test data) while detecting suicidal thoughts, which were academic performance, death of a loved one, psychological disorders including anxiety, family-related issues, past suicide attempts, other suicide attempts, and no suicide attempts, respectively, labeled as “label 0” through “label 5”, respectively (Figure 2). The results show that the model performs well in distinguishing between true positives and true negatives across the different labels, with high numbers of correctly identified cases in each matrix. For instance, in the confusion matrix for "label 0" (academic performance), the model correctly identified 272 instances as belonging to this category (while it incorrectly labeled 43 instances, a 86.35% accuracy), and 647 instances correctly identified as not belonging to it (while it incorrectly labeled 57 instances, a 91.9% accuracy). The false positive and false negative rates are relatively low, which indicates the model’s strong performance. Additionally, with a precision of 89%, a recall of 83%, and an F1-score of 86%, this model was highly reliable in identifying at-risk students (Table 1).

However, there are some variations in performance across different labels. For example, "label 5" (other suicide attempts) has a higher rate of false negatives (19 instances) as well as relatively lower precision, recall, and F1 scores compared to some of the other labels, which suggests that the model might occasionally miss identifying some cases in this category. While this method may not be as sophisticated in some areas compared to previous studies, it compensates by offering a thorough and contextually rich analysis of student-generated material.

Discussion

The fine-tuned model's performance, as reflected through accuracy, precision, and recall, indicates strong detection capabilities, especially in identifying at-risk students with minimal false positives and negatives. High recall in categories such as academic stress, death of a loved one, psychological disorders, and family-related issues suggests the model effectively captures these frequent mental health triggers; however, the confusion matrices show lower values of recall in broader categories such as “other suicide attempts,” most likely because of the uncertainty and overlap with more precise labels.

By emphasizing student-generated content and adopting multi-label classification to improve detection accuracy, this approach sets apart from other research. Unlike other models, it offers more focused and useful information by associating data with specific labels linked to frequent mental health issues among students. Integrating these results into a real-time monitoring system provides a valuable tool for early intervention and support. By precisely identifying at-risk individuals, the model can contribute significantly to timely interventions and support, which helps to address underlying mental health issues that broader approaches might overlook by focusing on subtle trends in student-written content.

Our approach has several limits even if it could identify mental health problems among students. For example, messages longer than 5,000 characters due to BERT's maximum token restriction could lead to lost important contextual information, which may be necessary for a complete understanding of the user's mental state; moreover, the reliance of this model on words associated with student life may overlook at-risk individuals whose discomfort signals do not fit these preset categories. Additionally, in order to enhance the performance for "secondary" and "other" labels, we could implement strategies such as expanding the training dataset to include more examples of less frequent categories, which would help the model learn their distinct features more effectively. Another strategy is to use advanced classification approaches such as hierarchical modeling or ensemble methods, which may help increase the model's ability to distinguish between overlapping groups (Liu et al., 2022).

Several methods would enhance the effectiveness and accessibility of the model for researchers wishing to expand on this method; first, expanding the dataset to incorporate more varied sources of text outside those especially connected to students could increase the generalizability of the model and assist in the identification of a wider spectrum of at-risk people. Also, incorporating multimodal data - such as physiological signals, audio, or video - may also offer a more encompassing picture of a person's mental state, which would circumvent some of the constraints of text-based study; in order to make sure that the labels more accurately reflect the meaning of the text, researchers may additionally investigate more complex data labeling methods, which include crowdsourcing or sophisticated natural language processing systems.

Acknowledgments

I thank Gubeom Kwon for giving code advice.

References

Aldhyani, Theyazn H H et al. "Detecting and Analyzing Suicidal Ideation on Social Media Using Deep Learning and Machine Learning Models." *International journal of environmental research and public health* vol. 19,19 12635. 3 Oct. 2022, doi:10.3390/ijerph191912635.

Allayla, Mohamed A., and Serkan Ayvaz. "A Big Data Analytics System for Predicting Suicidal Ideation in Real-Time Based on Social Media Streaming Data." *arXiv.Org*, 19 Mar. 2024, arxiv.org/abs/2404.12394.

Arowosegbe, Abayomi, and Tope Oyelade. "Application of Natural Language Processing (NLP) in Detecting and Preventing Suicide Ideation: A Systematic Review." *International journal of environmental research and public health* vol. 20,2 1514. 13 Jan. 2023, doi:10.3390/ijerph20021514.

Babu, Nirmal Varghese, and E Grace Mary Kanaga. "Sentiment Analysis in Social Media Data for Depression Detection Using Artificial Intelligence: A Review." *SN computer science* vol. 3,1 (2022): 74. doi:10.1007/s42979-021-00958-1.

Bilsen, Johan. "Suicide and Youth: Risk Factors." *Frontiers in psychiatry* vol. 9 540. 30 Oct. 2018, doi:10.3389/fpsyt.2018.00540.

Deshpande, Mandar, and Vignesh Rao. “Depression Detection Using Emotion Artificial Intelligence.” *International Conference on Intelligent Sustainable Systems (ICISS)*, 21 June 2018, doi:10.1109/ISS1.2017.8389299.

Devlin, Jacob, et al. “Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” *arXiv.Org*, 24 May 2019, arxiv.org/abs/2405.05795.

Diniz, Evandro J S et al. “Boamente: A Natural Language Processing-Based Digital Phenotyping Tool for Smart Monitoring of Suicidal Ideation.” *Healthcare (Basel, Switzerland)* vol. 10,4 698. 8 Apr. 2022, doi:10.3390/healthcare10040698.

Grimmond, Jessica et al. “A qualitative systematic review of experiences and perceptions of youth suicide.” *PloS one* vol. 14,6 e0217568. 12 Jun. 2019, doi:10.1371/journal.pone.0217568.

Liu, Lijue et al. “Solving the class imbalance problem using ensemble algorithm: application of screening for aortic dissection.” *BMC medical informatics and decision making* vol. 22,1 82. 28 Mar. 2022, doi:10.1186/s12911-022-01821-w.

Orozco, Ricardo et al. “Association between attempted suicide and academic performance indicators among middle and high school students in Mexico: results from a national survey.” *Child and adolescent psychiatry and mental health* vol. 12 9. 24 Jan. 2018, doi:10.1186/s13034-018-0215-6.

Ravindranath, Orma et al. “Adolescent neurocognitive development and decision-making abilities regarding gender-affirming care.” *Developmental cognitive neuroscience* vol. 67 (2024): 101351. doi:10.1016/j.dcn.2024.101351.

Rogers, Anna, et al. “A Primer in BERTology: What We Know about How BERT Works.” *arXiv.Org*, 9 Nov. 2020, arxiv.org/abs/2002.12327.

Thakkar, Anoushka et al. “Artificial intelligence in positive mental health: a narrative review.” *Frontiers in digital health* vol. 6 1280235. 18 Mar. 2024, doi:10.3389/fdgth.2024.1280235.

Van Der Lee, M.M.E. “Suicide Severity Risk Prediction Using BERT.” *Tilburg University*, 24 June 2022, arno.uvt.nl/show.cgi?fid=161190.