Mental Health in the Digital Era-NLP Models for Depression and Suicidal Tendency Detection

Rohan C[1] and Dr. Sapna VM[1]

PES University, Bengaluru KA 560085, India,
chandrashekar.rohans@gmail.com,
sapnavm@pes.edu

Abstract. The advent of Natural Language Processing (NLP) as a novel technology, especially with the integration of the BERT (Bidirectional Encoder Representations from Transformers) model, presents a transformative approach in addressing mental health issues in today's digital landscape. BERT's advanced capabilities in understanding the context and nuances of language enable more accurate detection of mental health concerns, including depression and suicidal tendencies, in user-generated text. This exploration presents an in-depth exploration of a cutting-edge NLP-based model designed to detect signs of mental health issues, depression, and suicidal tendencies in user-generated text. The authors emphasise the importance of leveraging the advancements in AI and machine learning for early detection of mental health issues. We introduce a novel approach using NLP models, distinguishing it from the traditional machine learning models that are commonly employed. The end goal is to facilitate better intervention methods, offering a more comprehensive view of an individual's mental well-being, especially when a vast majority of communication has shifted to digital platforms.

Keywords: mental health, natural language processing, suicidal tendencies, computational linguistics, contextual language processing

1 Introduction

Mental health, an integral component of overall well-being, has garnered increasing attention, especially in the digital era. The proliferation of digital platforms for communication has opened new channels for understanding and addressing mental health issues. Social media and online forums, where individuals express their emotions and experiences, provide a rich source of data for detecting mental health concerns such as depression and suicidal tendencies.

Depression, a major contributor to the global burden of disease, affects millions worldwide. This underscores the need for effective detection and intervention strategies. Traditional methods of mental health assessment, based primarily on patient-clinician interactions and self-reporting, face limitations in scalability and timeliness, often compounded by stigma associated with mental health conditions.

In response to these challenges, Natural Language Processing (NLP) offers a promising solution. NLP, which lies at the intersection of computer science, artificial intelligence, and linguistics, has seen significant advancements due to breakthroughs in machine learning and deep learning. These technologies enable the nuanced processing and understanding of human language, making NLP a powerful tool for mental health applications.

Our research focuses on developing an NLP-based model for detecting signs of depression and suicidal tendencies in user-generated text. This approach leverages the advancements in AI and machine learning to analyze textual data on digital platforms, identifying patterns indicative of mental distress. As highlighted by Haidt and Sharma et al. ([1] and [2]), and other studies, digital communication provides valuable insights into an individual's mental state, facilitating early intervention.

This paper will explore the unique aspects of our NLP approach, its technical foundations, and the implications of our findings for mental health care in the digital era. We aim to contribute to the evolving landscape of mental health interventions, providing new tools and perspectives for tackling mental health challenges in today's digitally interconnected world.

2 Novelty

In the pursuit of innovating mental health diagnostics, the incorporation of Natural Language Processing (NLP) marks a significant shift from traditional methodologies. Unlike conventional approaches that often rely on overt, explicit expressions in text, NLP enables a deeper, more nuanced analysis. This novel approach recognizes the complexity of human language and emotional expression, particularly in online communications.

This not only enhances the accuracy of detection but also broadens the scope of what can be discerned from digital communications. The adoption of this technology represents a paradigm shift in mental health diagnostics, promising more effective, timely, and empathetic understanding of individuals' mental states, a critical advancement in the realm of digital-era healthcare.

2.1 Traditional vs NLP Approaches

The assessment of suicide risk through online postings has been explored by Saab et al.[3], highlighting the power of computational models in this domain. The combination of ubiquitous sensors and machine learning offers a promising avenue for understanding mental health, as discussed by Castillo et al.[4].

However, traditional methods in the realm of mental health detection through textual data have primarily relied on keyword-based approaches or basic pattern recognition. These methods would identify specific words or phrases commonly associated with mental distress or depressive sentiments, and then categorize the text based on the presence or absence of these indicators. For instance, detecting words such as "sad", "lonely", or "hopeless" might flag a piece of text as potentially indicative of depression.

2.2 Limitations of Traditional Techniques

Lack of Contextual Understanding: Traditional techniques often failed to understand the nuanced and contextual nature of language. A statement like "I'm feeling blue today because of the weather" might be inaccurately flagged due to the presence of the word "blue", even though the sentiment is lighthearted.

Over-Reliance on Keywords: Merely depending on a set of predefined keywords could lead to a significant number of false positives or negatives. Language, especially in the context of mental health, is deeply personal and can vary widely among individuals.

Inability to Detect Implicit Expressions: Often, individuals might express their feelings or distress implicitly, without using overtly negative words. Traditional methods could miss out on these subtle indicators.

2.3 Uniqueness of The NLP Approach

The NLP-based approach, as employed in this research, addresses the aforementioned limitations by offering a more in-depth, contextual understanding of textual data. NLP models, especially advanced ones like BERT, are trained on vast amounts of text, enabling them to grasp the intricate nuances and semantics of language.

Contextual Understanding: BERT, which stands for Bidirectional Encoder Representations from Transformers, is particularly known for its bidirectional understanding of text. This means it doesn't just look at words in isolation but considers the entire sentence or even paragraph to infer meaning.

Reduction in False Positives: With a richer understanding of context, the NLP approach can significantly reduce false positives, as highlighted in the comparison metrics.

Detection of Implicit Expressions: Advanced NLP models can pick up on patterns that might not be immediately obvious, recognizing implicit signs of distress or mental health concerns that traditional methods might overlook.

In essence, the NLP-based model introduced in this research not only offers superior performance metrics but also provides a more comprehensive, nuanced, and human-like understanding of textual data, making it a powerful tool in the early detection of mental health issues in the digital era.

3 Dataset

This section delves into the technical aspects of our dataset, which forms the backbone of our NLP model's effectiveness. We carefully detail the process of selecting, preparing, and structuring our dataset to ensure it's well-suited for identifying mental health indicators like depression and suicidal tendencies. The accuracy and reliability of our model largely depend on the quality and composition of this data. We'll explain how we balanced the dataset and the rationale behind our choice of data sources, providing a clear, straightforward insight into these critical preparatory steps. This focus on dataset integrity is essential for the success of our NLP application in mental health analysis.

3.1 Description and Preparation

The dataset utilized in this research is a collection of 27,977 text entries. A sample of which can be seen in Table 1. These text entries are from a variety of social media platforms mainly Twitter and Reddit. Liu et al. [5] showcased techniques to recognize signs of depression on Twitter, further solidifying the platform's utility in mental health research. Kim et al.[6] pioneered efforts in predicting depression using social media data, emphasizing the potential of such platforms. Platforms like Instagram have also been used to identify markers of depression, as discussed by Reece et al.[7]. Each entry in the dataset corresponds to a message or post from individuals discussing a range of topics, primarily centered around mental health, personal experiences, and related subjects. The dataset is structured with two primary columns:

Text: This column encapsulates the actual content of the message. The text entries vary in length and detail but provide a rich source of information that is crucial for our analysis.

Label: The dataset employs a binary labeling system. The label 0 signifies messages that are neutral or non-critical in nature. On the other hand, a label of 1 is used to tag messages that indicate potential mental health distress or concern. This labeling enables us to classify and analyze the severity or criticality of the messages effectively. Sarsam et al.[8] delved into the machine classification of suicide-related communication on platforms like Twitter.

In terms of distribution, the dataset is fairly balanced. Approximately 50.54% of the messages are labeled as 0, while 49.46% carry a label of 1. This balanced distribution is beneficial for our analysis as it prevents any significant bias towards a particular label, ensuring that our models and algorithms operate on a representative sample of data.

3.2 The BERT Model

BERT, which stands for Bidirectional Encoder Representations from Transformers, is a state-of-the-art language representation model (as shown in Fig. 1). In-

Table 1. A sample table illustrating the shape of the dataset.

Text (First 50 Characters)	Label
dear american teens question dutch person heard gu...	0
nothing look forward lifei dont many reasons keep ...	1
music recommendations im looking expand playlist u...	0
im done trying feel betterthe reason im still aliv...	1
worried year old girl subject domestic physicalme...	1

troduced by Kenton et al.[9], BERT has transformed the landscape of natural language processing with its deep bidirectional transformers. It has revolutionized the field of (NLP) due to its superior ability to understand the context of words in a sentence. Instead of looking at words in isolation, BERT examines words concerning their neighbors in both directions, leading to a much richer representation.

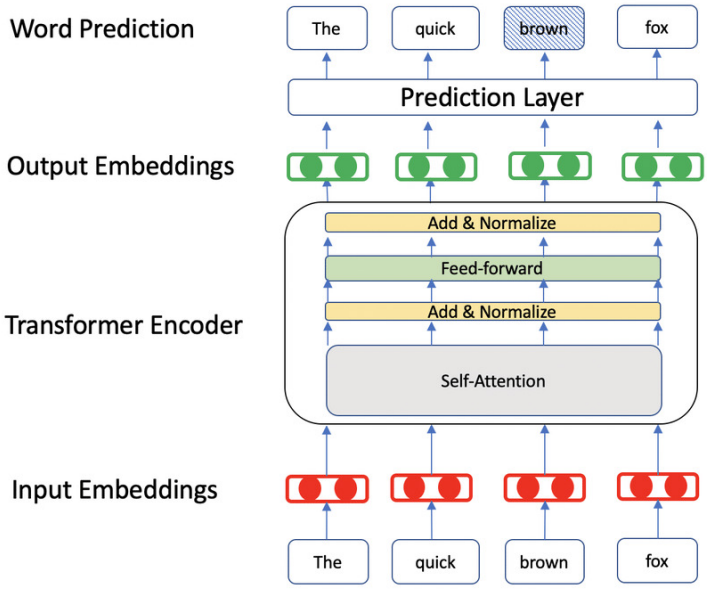

Fig. 1. An illustration of the BERT model.

Key features of BERT include:

Bidirectionality: Traditional language models, like LSTM or GRU, operate either from left-to-right or right-to-left. In contrast, BERT processes words con-

cerning all other words in the sentence, making it genuinely bidirectional. This feature allows BERT to capture intricate nuances and relationships between words that other models might miss.

Transformers: BERT is built upon the Transformer architecture, which employs self-attention mechanisms to weigh the significance of words in a sentence. This architecture enables BERT to focus more on words that are more relevant in a given context.

Pre-training on vast datasets: BERT is pre-trained on massive amounts of text, including the entire Wikipedia and the BookCorpus dataset. This extensive pre-training allows BERT to develop a comprehensive understanding of language.

Fine-tuning for specific tasks: While BERT is pre-trained on general data, it can be fine-tuned on specific datasets to perform a wide range of tasks, from text classification to question-answering. This adaptability makes BERT incredibly versatile and suitable for diverse NLP challenges. Also, the concept of attention mechanisms, which has been foundational for models like BERT, was detailed by Clark et al.[10].

For our research, BERT offers several advantages. Its deep understanding of context makes it particularly adept at discerning subtle indications of mental health distress in text messages. Furthermore, BERT's capability to capture long-range dependencies in texts helps in analyzing longer messages where crucial information might be spread out.

4 Results and Comparison

In this section, we present the results obtained from our experiments, along with comparisons to benchmark models and previous research

4.1 Experimental Setup

All experiments were conducted under a consistent environment to ensure fairness in comparison. We employed a train-test split to segregate our dataset, with 80% used for training and the remaining 20% for testing. The models were trained using a consistent set of hyperparameters, and the results were averaged over multiple runs to account for any variability.

4.2 Model Performance

Our primary model, BERT, exhibited promising results. Its capability to understand context and capture intricate relationships in the text was evident from its

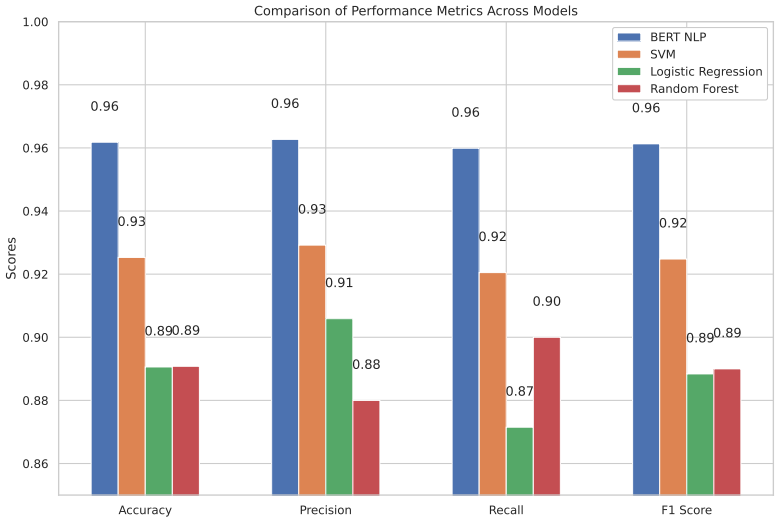

Fig. 2. This comparison plot illustrate the effectiveness of our NLP-Based approach by showcasing several key performance metrics, including accuracy, precision, recall, and F1 -score.

high accuracy and F1-score (as shown in Fig. 2). The confusion matrix revealed that BERT was particularly adept at minimizing false negatives, which is crucial given the sensitive nature of our task.

In comparison, traditional models like LSTM and GRU performed reasonably but lagged behind BERT in terms of precision and recall. This disparity underscores the importance of bidirectional context understanding, a feature inherent to BERT.

4.3 Comparison with existing work

Compared to previous research in this domain, our approach stands out in its use of advanced NLP models and a balanced dataset. Earlier studies have often shown to have failed to understand the context of the user's text and as a result misinterpret their mental health status. Bayesian techniques, as explored by Ji et al.[11], have also been employed to discover patterns of mental disorders in online forums. While acknowledging that previous research work has provided great insights. The NLP-BERT model seems to have an indisputable edge over the traditional Machine Learning (ML) models.

5 User Interface and real-time application

To make the model accessible and user-friendly, a rudimentary GUI has been designed. Users can input a sample text, which is then analyzed to provide a risk score (refer to Fig. 3 and Fig. 4). The model also checks for specific trigger words, flagging them as potential warning signs. This instant feedback can be immensely useful for mental health professionals, allowing them to understand their patients' state rapidly.

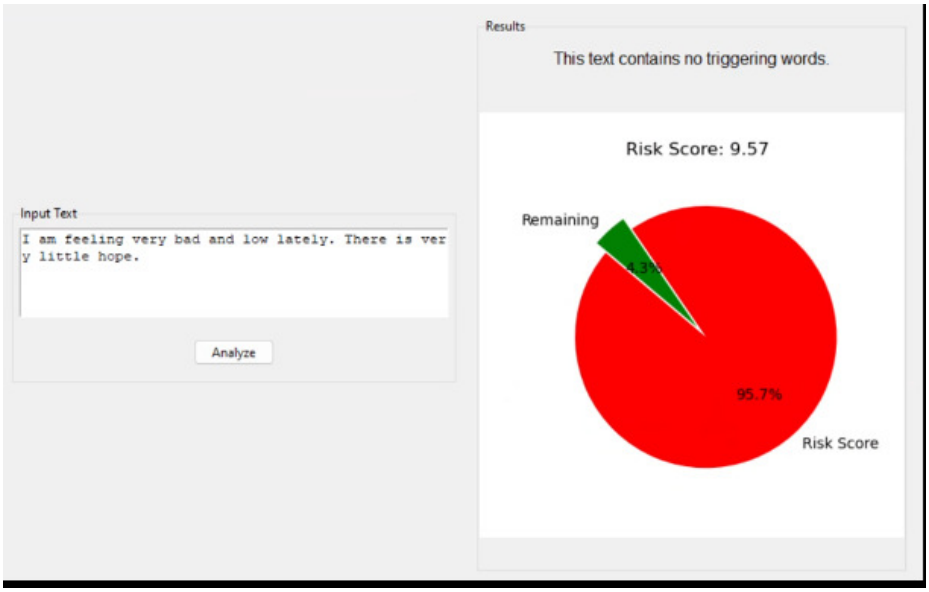

Fig. 3. GUI displaying a text of a user as "At Risk". The model is able to identify that the user's text shows signs of mental health distress.

6 Future scope and potential

The field of mental health diagnosis is always changing and we're about to see some big improvements. This new model is a great example of the exciting changes coming up. As we plan for the future, there are many features we can add to make our current system even better.

One of the key changes is including a person's past messages and when they sent them. This can give us a better understanding when and why a person's feelings might change. By looking at past messages, we can see patterns over long periods of time. This gives us a better picture of someone's emotional changes and helps us notice things that might be easy to miss otherwise.

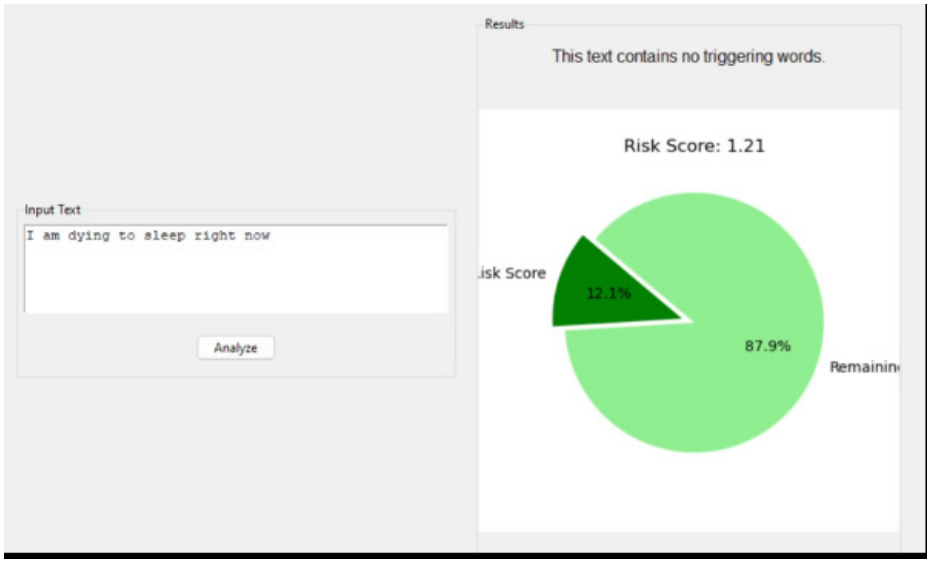

Fig. 4. GUI Displaying text of a user "Not at Risk". The model can clearly discern that even though the word 'dying' is present, when viewed in context, it does not show signs of mental distress.

In addition, there's a lot of data that we haven't used yet which can be very helpful. Forums have been another medium where NLP models, as mentioned by Shrestha et al.[12], have been applied to assess depression and self-harm risks. By using larger datasets, our model can become more versatile. Parapar et al.[13] discussed the importance of early risk prediction on the internet, emphasizing the need for timely interventions. This can lead to offering help sooner and supporting people in a timely way.

7 Concluding Remarks

In our work to improve mental health diagnosis, our new NLP model is a big step forward. However, it's important to see it as a part of a bigger picture in mental health care. Even though our model shows how far we have come in using technology for mental health, it can't be the only solution we rely on. We must also keep in mind the ethical considerations for using social media data as described by Garg and Dhiman et al.([14]and [15]).

The real value of our model is twofold: it's a sign of new tech advances and it works alongside traditional therapy methods. It can be used in real-time to give therapists, caregivers, and even police important information. However, it's up to us, the people using this technology, to use it wisely. It's important to know what the model can and can't do. While it can give advice and share information, it can't replace the deep understanding and care that humans provide.

In the end, our NLP model shows what the future might look like, but it also reminds us how important human connection is in mental health care. As we move ahead, our model should work alongside human therapists, not replace them, adding to the care and understanding they provide.

References

1. Haidt, J. and Allen, N., 2020. Scrutinizing the effects of digital technology on mental health.
2. Sharma, M.K., John, N. and Sahu, M., 2020. Influence of social media on mental health: a systematic review. Current opinion in psychiatry, 33(5), pp.467-475.
3. Saab, M.M., Murphy, M., Meehan, E., Dillon, C.B., O'Connell, S., Hegarty, J., Heffernan, S., Greaney, S., Kilty, C., Goodwin, J. and Hartigan, I., 2022. Suicide and self-harm risk assessment: a systematic review of prospective research. Archives of suicide research, 26(4), pp.1645-1665.
4. Castillo-Sánchez, G., Marques, G., Dorronzoro, E., Rivera-Romero, O., Franco-Martín, M. and De la Torre-Díez, I., 2020. Suicide risk assessment using machine learning and social networks: a scoping review. Journal of medical systems, 44(12), p.205.
5. Liu, D., Feng, X.L., Ahmed, F., Shahid, M. and Guo, J., 2022. Detecting and measuring depression on social media using a machine learning approach: systematic review. JMIR Mental Health, 9(3), p.e27244.
6. Kim, J., Lee, J., Park, E. and Han, J., 2020. A deep learning model for detecting mental illness from user content on social media. Scientific reports, 10(1), p.11846.
7. Reece, A.G. and Danforth, C.M., 2017. Instagram photos reveal predictive markers of depression. EPJ Data Science, 6(1), p.15.
8. Sarsam, S.M., Al-Samarraie, H., Alzahrani, A.I., Alnumay, W. and Smith, A.P., 2021. A lexicon-based approach to detecting suicide-related messages on Twitter. Biomedical Signal Processing and Control, 65, p.102355.
9. Kenton, J.D.M.W.C. and Toutanova, L.K., 2019, June. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of naacL-HLT (Vol. 1, p. 2).
10. Clark, K., Khandelwal, U., Levy, O. and Manning, C.D., 2019. What does bert look at? an analysis of bert's attention. arXiv preprint arXiv:1906.04341.
11. Ji, M., Xie, W., Zhao, M., Qian, X., Chow, C.Y., Lam, K.Y., Yan, J. and Hao, T., 2022. Probabilistic Prediction of Nonadherence to Psychiatric Disorder Medication from Mental Health Forum Data: Developing and Validating Bayesian Machine Learning Classifiers. Computational Intelligence and Neuroscience, 2022.
12. Shrestha, A., Serra, E. and Spezzano, F., 2020. Multi-modal social and psycho-linguistic embedding via recurrent neural networks to identify depressed users in online forums. Network Modeling Analysis in Health Informatics and Bioinformatics, 9, pp.1-11.
13. Parapar, J., Martín-Rodilla, P., Losada, D.E. and Crestani, F., 2023, September. Overview of erisk 2023: Early risk prediction on the internet. In International Conference of the Cross-Language Evaluation Forum for European Languages (pp. 294-315). Cham: Springer Nature Switzerland.
14. Garg, M., 2023. Mental health analysis in social media posts: a survey. Archives of Computational Methods in Engineering, 30(3), pp.1819-1842.
15. Dhiman, D.B., 2023. Ethical Issues and Challenges in Social Media: A Current Scenario. Available at SSRN 4406610.