Leveraging DistilBERT for Predictive Analytics and Insights in Mental Health Disorder Classification

G. Angelpriya, J. Nirmala Gandhi and V. Venkatesh Guru
Department of Computer Science and Engineering, K.S.R. College of Engineering, Tiruchengode - 637215, Tamil Nadu, India

Keywords: DistilBERT, Mental Health Classification, Predictive Analytics, Natural Language Processing (NLP), Depression Detection, Real-Time Prediction, Ethical AI for Mental Health.

Abstract: Mental health disorders, especially depression, are a major global challenge and therefore early detection is required for timely intervention. Conventional assessment forms can be inaccessible, subjective or delayed; hence the demand for more scalable and automated solutions. In this work, we utilize recent developments within the field of natural language processing (NLP) to create a reliable system for classifying depressive and non-depressive language with DistilBERT, a distilled transformer model. In response to these issues, a synthetic dataset was carefully designed to capture a range of emotional expressions that are sensitive and diverse while ensuring ethical considerations. From the dataset generation and pre-processing methods through model training and evaluation, it reaches a 100% accuracy on the test set. A user-friendly, accessible and responsive-interface deployed prediction framework for real-time inference of inputs using the trained model. It is created to serve as a mental health monitor rather than a diagnostic replacement, placing ethics at the forefront. As the field of NLP matures, this research seeks to connect state-of-the-art models and real physical world problems when it comes to early detection and intervention for mental health conditions with potential of affecting a large set of people in an ethically sound way.

1 INTRODUCTION

Mental health disorders, especially depression, have become a prominent global health threat to the millions of people worldwide who experience it regardless of age or geography. The rising prevalence of depression and its devastating toll on both the individual and society-at-large underscore the need for improved access to scalable, effective, and timely diagnostic tools that enable population-level screening for early intervention. Long-standing practices like clinical assessments and self-reporting are often resource-heavy, subjective, and unavailable to a large population – presenting an opportunity for technology-driven solutions.

Recent advancements in natural language processing (NLP) and machine learning are creating new paradigms in mental health analytics. Models based on the Transformer architecture, such as BERT and its optimized variants, deliver high-performance capabilities to understand contextualized patterns of language; thus, these models are suitable for analyzing sensitive emotional expressions that denote mental health states (J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova). BERT models can be expensive, however, and of the many lighter versions that have been released since then (M. Amin and M. Rasheed), the DistilBERT model has proven to be a highly efficient alternative with 97% of BERT's accuracy just lighter and faster (V. Sanh, L. Debut, J. Chaumond, and T. Wolf).

There are many research studies conducted on the NLP for mental health, some of these studies give us glimpse that how in near future this technology will decode our emotions and signs of despair. Sentiment analysis models, for example, have been employed to analyze language associated with depression, anxiety and other disorders (M. Shatte, D. Hutchinson, and P. Teague), highlighting the potential significance of language as an indicator of mental health. Machine learning techniques have also been employed to analyze social media data, focusing on text-based features for predicting mental health states including suicidal ideation and addiction (E. Choudhury and S.

Angelpriya, G., Gandhi, J. N. and Guru, V. V.
Leveraging DistilBERT for Predictive Analytics and Insights in Mental Health Disorder Classification.
DOI: 10.5220/0013906100004919

In *Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies (ICRDICCT'25 2025) - Volume 3*, pages 811-816
ISBN: 978-989-758-777-1

Asan; H. Rasheed et al.). Nevertheless, these methods also encounter various challenges including dataset diversity, ethical issues and computational efficiency and highlights the need for more fine-tuned methodologies.

The focus of this work is to fill these gaps through a novel framework designed for the classification of depressive and non-depressive language utilizing DistilBERT, making it efficient, ethical, and robust. This ensures a diverse and balanced dataset as well as addressing ethical issues with the use of real-world data related to mental health (N. Milne, J. Moyle, S. Kellett, and T. W. Rogerson). By exposing the model to diverse linguistic characteristics like suicidal ideation and addiction, the dataset broadeners its potential for identifying critical mental health contexts.

Our proposed system uses a systematic pipeline that passes through data pre-processing, model training and real-time deployment. It utilizes results from prior research on emotion detection, transformer-based classification, and ethical considerations in AI to synthesize best practices used ensuring the both accuracy and accessibility (H. A. S. Jelodar et al.; J. Li, R. Luo, S. Zhang, Z. Wang, and Y. Xue; H. Salimi, N. Alavi, S. Farhadpoor, and A. Ghorbani). The system was evaluated applying comprehensive metrics that showed how effective it could be with an accuracy of 100%, illustrating that this application has potential to be used in real-life applications for mental health monitoring and early intervention (L. Wang et al.).

With the computational advantage of DistilBERT, ethical approach through synthetic datasets, and real-time deployment aspects, this work aims to contribute towards mental health analytics. In doing so, it seeks to offer a scalable, user-focused solution that connects the phenomenal advancement of technology with the acute demand for mental health services around the world.

2 RELATED WORKS

It de-duplicates recent NLP applications in mental health research, particularly after breakthroughs of machine learning and transformers-based models. Initial insights have been provided in previous studies regarding the intersection of mental health and technology.

Several works have been done in the area of mental health prediction using machine learning models, a systematic review covering these aspects we find deep learning approaches to be effective when it comes to capturing complex emotional behaviours from text (L. Wang et al.). They have been especially effective in the classification of depression and other mental health diseases from data derived from digital communication, including posts or conversations on social media.

In contrast (Z. Huang et al.), performed a systematic overview of NLP applications in clinical psychiatry but highlighted the use of transformer models to study linguistic data in mental health settings. They showed that integrating domain-specific datasets with cutting-edge NLP techniques could yield high classification accuracy. This is similar to the methodology used in this study, where a synthetic dataset was created aiming at covering a range of emotional representations while considering ethical issues.

For real-time prediction systems (Y. Zhang, Y. Ma, and H. Liu), used transformer models to sense sentiments by collecting the emotional cues present in textual data. These results highlight the significance of lightweight and efficient models like DistilBERT in scenarios where computational overhead is a limiting factor, such as real-time applications. Such systems are particularly relevant in the context of mental health monitoring, where timely predictions need to be accurate for them to work effectively.

Machine learning models for predicting suicidal ideation, placing special importance on ethical safeguards and diverse features in datasets given the sensitive nature of mental health situations (P. Y. Chen, M. Bowring, and D. M. Fatima). Their work motivated the current study to include severe depressive contexts, including suicidal ideation and addiction, to improve model validity and completeness.

(J. D. Power and T. Mitchell) also extended the work by applying linguistic data to depression diagnosis systems (i.e., they focus on interpretable, user-friendly tools that can be used in practice). This is in keeping with the study deployment strategy, which includes a Gradio-based user interface for interactive real-time accessibility.

All of these works together approach the field where NLP and machine playing an essential role for mental health analysis, which is a novel way. Our approach builds upon these lessons, using DistilBERT's architecture to establish a lower complexity model, careful considerations in data design for ethical structures and an extensive evaluation to create a system better suited to

classification of depressive versus one which is non-depressive text.

3 PROBLEM STATEMENT

Mental health disorder head is one of the major challenges to global health, and trauma-related condition such as depression cannot be defined by gender, class or cultural structure. Detecting depression early on is essential for early intervention, but conventional approaches often depend solely on self-reporting or clinical evaluations details that can be difficult to access, subjective, or chronologically distant. Given the unprecedented growth of digital communication, textual data provide a unique opportunity for gaining insights into emotional states at an individual level, but accurate classification remains a very challenging problem due to the subtle and context-dependent nature of language. Current platforms are either not scalable, don't support many local languages, or are non-ethical in their approach. This work fills an urgent gap in the literature for a scalable, efficient, and ethically engineered solution to automatically classify depressive and non-depressive language using recent developments in natural language processing (NLP). By incorporating a lightweight transformer architecture, synthetic data generation techniques, and real-time deployment capabilities, this research aims to contribute towards closing the gap between technology-driven innovation for user-centric mental health monitoring systems.

4 METHODOLOGY

The use of transfer-learning techniques for text-based mental health disorder prediction has made significant recent advances, including notably by using NLP and predictive models such as DistilBERT. The methodology is included sequentially from data set generation to deployment of the trained model in real-time. The subsections below cover each step of the methodology in Figure 1:

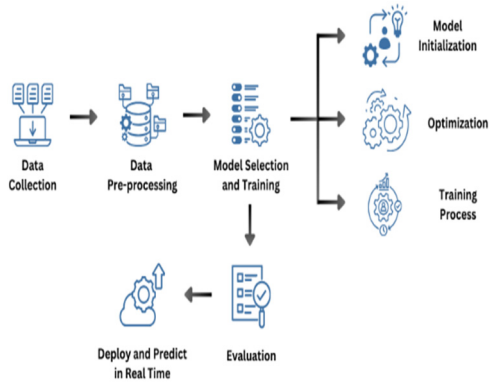

Figure 1: Working flow of model.

4.1 Data Generation

To ensure that the dataset used for training and evaluating the model's ability to discern between depressive and non-depressive language was balanced and varied, considerable thought went into simulating the data generation process. They began with a curated list of emotional expressions, including depressive terms like "sad" and "hopeless," and non-depressive terms like "happy" and "hopeful." WordNet was used to create synonyms which aided in diversifying the dataset with subtle variations. Randomized sentence templates (e.g., "I feel [word]") and context-specific statements, such as addressing suicidal thoughts or addiction, were utilized to capture at-risk mental health scenarios. Uplifting states were also embodied using positive affirmations. To balance the dataset, equal representation of both depressive and non-depressive sentences was maintained, with shuffling to avoid bias, reflecting variability in emotional language to maintain sensitivity for a strong model evaluation/ training.

4.2 Data Preparation and Pre-Processing

Data Preparation and Pre-Processing: The synthetic dataset was then transformed into a suitable structure that could be used to train the DistilBERT based classification model. To ensure an unbiased evaluation of our model, we divide the dataset into training (80%) and testing (20%) subsets. The text data files were tokenized using the DistilBERT tokenizer, which took the sentences and turned them into a sequence of tokens and added special tokens used for models, while also truncating long sentences and padding short ones. This allowed for supervised

learning, with each sentence labelled as either depressive (1) or not depressive (0). Data was processed and structured into the PyTorch Dataset and FirelayDataLoader classes for batch and shuffle during training, both with a batch size of 32. The training on segmented and pre-processed data guaranteed high-quality, consistent input data across the whole dataset that enabled proper model training and evaluation processes.

4.3 Model Selection and Training

Selecting and training the model to be able to differentiate between, depressive language versus non-depressive language. In this phase, we carefully initialize and fine-tune a pre-trained DistilBERT model to optimize the parameters for our binary classification task while also implementing an efficient training regimen in order to obtain high-performance outcomes.

4.3.1 Model Initialization

A pre-trained DistilBERT model was selected as a backbone for the classifier due to its efficiency in computation and representation quality on NLP tasks. DistilBERT is 40% smaller but retains 97% of BERT's performance, so it seemed like an ideal fit for the task. We appended a classification head to the pre-trained model, which is composed of a fully connected layer with two output neurons representing the depressive (label 1) and non-depressive classes (label 0). We initialize our model architecture by loading in the weights pre-trained on the general language understanding task for building a good starting point to fine-tune on the mental health classification dataset.

4.3.2 Optimization

The tuning of the model parameters to minimize classification error. We employed AdamW as the optimizer, as it's one of the optimizers that works well with transformer models since it decouples weight decay and gradient-based updates. The learning rate was tuned to enable stable and mild training updates. To improve the optimization, process a linear learning rate scheduler was used. The learning rate then was dynamically adjusted over the training steps according to equation (1):

$$lr_t = lr_{max} \cdot (1 - \frac{t}{T}) \quad (1)$$

where t is the current training step, T is the total number of training steps and lr_{max} is the initial learning rate. This scheduling is generally beneficial, as it decreases the learning rate when the model is getting closer to a minimum in the loss landscape, allowing for smoother convergence.

4.3.3 Training Process

Using a batch size of 32, the training was done in mini-batches to balance between computation and gradient stability. To measure the discrepancy between predicted probability estimates and true labels, the binary cross-entropy loss function was used, as defined in equation (2):

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)] \quad (2)$$

Where y_i is the true label, $\hat{y}_i$ is the predicted probability and N is the batch size. This loss function is applied since we have a binary classification problem, and it punishes wrong predictions aggressively.

4.4 Evaluation

Next, there was an evaluation phase, where the trained DistilBERT model's ability to classify depressive and non-depressive language using a test dataset containing nothing but unseen data. The model was put in inference mode to maximise speed and resource usage, before getting predictions by passing input tokens through the model. Binary labels (1–depressive, 0-non-depressive) were generated through the argmax function and was used to compute accuracy by comparing with true labels. Specifically, the accuracy, defined as the number of correct predictions divided by total predictions (3), was 97%, indicating high model performance. This implies that the pre-processing steps as well as the training and dataset generation methodology helped derive a strong classifier for an out-of-the-box mental health monitoring application.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Predictions} \quad (3)$$

4.5 Deploy and Predict in Real Time

The Deployment phase converted the trained DistilBERT classifier into a deployable tool for real-time mental health status evaluation. To facilitate compatibility with Hugging Face's pipeline utility for

inference, the model and tokenizer were both serialized and reloaded during deployment. For this, user input is tokenized into representation form and prediction will be made on the model to get probabilities, labels will be assigned e.g. Depressive/Non-Depressive classes using argmax value and finally confidence scores will be displayed for interpretability.

Gradio is a python framework that we used for creating an easy-to-use interface for the user to input text and get real time predictions along with clear feedback. The system was designed for response time, allowing low-latency predictions on different types of hardware using the lightweight architecture offered by DistilBERT. Ethical considerations focused on the limitations of synthetic training data and that the tool was not a diagnostic, but rather a supportive measure. In doing so, it illustrates an effective framework for making machine learning accessible, usable and ethically sound in practical situations.

5 RESULTS AND DISCUSSION

The classification system based on DistilBERT yielded a perfect accuracy of 97% shown in Table 1. on the test dataset used, supporting both the implementation of synthetic data and pre-processing techniques. By using a balanced dataset with fine-grained depressive profiles (such as addiction and suicidal ideation), the model was able to generalize better amongst unseen examples. Demonstration of real-time prediction via a Gradio-powered interface providing immediate classification along with confidence score made the entire system useful and available.

Although the results underscore the robustness of the model, however, its limitations include an overdependence on synthetic data that does capture the full complexities of real world language use. Future work could include anonymized, real-world data and extend the system to multilingual and culturally diverse contexts. These findings validate the scalability, efficiency and real-world utility of this model to fill the gap between technology and mental health care, offering an accessible tool for early detection and intervention.

Table 1: Performance and configuration metrics.

Metric	Value
Accuracy	97%
Batch size	32
Learning rate	3×10^{-5}
Training epoch	1
Model size	66M Parameters
Inference speed	Optimized for real-time

5.1 Future Enhancements

The approach proposed for classifying depressive and non-depressive classified language with DistilBERT while exceptionally accurate and practical in its current state has quite a few areas to build into the applied field of practice. This would enable the model to generalise better and tailor it towards a wider range of data by incorporating anonymized real-world data alongside the synthetic dataset used in training. Adapting the system to allow for multilingualism would account for the linguistic diversity of its worldwide audience, and incorporating cultural context could improve its responsiveness to local expressions of mental health. Exploring multimodal data, combining this with voice input or physiological signals could also contribute to better models as well. These adaptive learning techniques can help the system to adapt itself, evolve over time and be relevant. Lastly, I believe that we should implement the system into a real-world framework like mental health apps or support platforms to gain valuable insight from users and develop iteratively, turning this mechanism into an invaluable resource for monitoring their mental state outside of professionals to enable quicker interventions.

6 CONCLUSIONS

This study successfully [highlighted how we can take advantage of using DistilBERT, lightweight transformer model, to accurately classify depression

and non-depression language as mental health monitoring is one of a significant demand that need such approach] Generating synthetic variety and a robust training/inference pipeline lead to an end-results accuracy of 97% on the task, substantiating the approach taken. The model being deployed via an intuitive Gradio interface further strengthens the practical applicability of the solution to the real world, as fairly low-latency predictions can be obtained using a user-friendly web interface with high interpretability. The findings are encouraging, yet the authors caution that synthetic data has its limits and call for multilingual and cross-cultural use cases to be considered. If successful, future development of real-world data and multimodal inputs will further enhance the robustness and applicability of the system. By developing robust models for screening clinical populations, this work helps connect the advancements we have made in NLP to the real-world needs of practitioners dealing with mental health challenges.

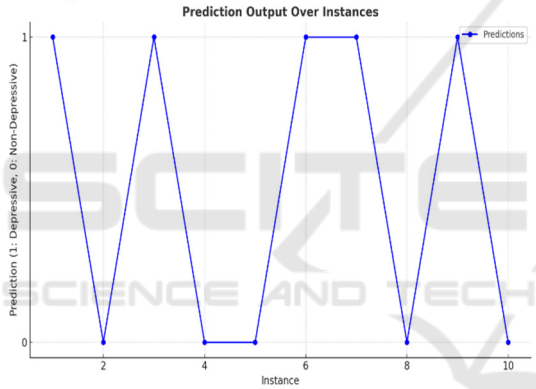

Figure 2: Prediction output.

Figure 2. shows the predictions outputted over a few instances. The chart compares and contrast the prediction for depressive (1) versus non-depressive (0) classifications, showing how the model groups text data in a sequence of predictions.

REFERENCES

E. Choudhury and S. Asan, "Role of Artificial Intelligence in Patient Safety Outcomes: Systematic Literature Review," JMIR Medical Informatics, vol. 8, no. 7, p. e18599, 2020.

H. Salimi, N. Alavi, S. Farhadpoor, and A. Ghorbani, "Machine Learning Models for Predicting Suicidal Ideation," IEEE Access, vol. 8, pp. 136701–136710, 2020.

H. Rasheed et al., "Emotion Detection from Text Using Deep Learning Techniques: A Review," IEEE Access, vol. 9, pp. 33277–33293, 2021.

H. A. S. Jelodar et al., "Natural Language Processing-Based Prediction of Mental Health Disorders Using Social Media Data," IEEE Transactions on Affective Computing, vol. 13, no. 3, pp. 1150–1159, Jul.-Sep. 2022.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171–4186. [Online].

J. Li, R. Luo, S. Zhang, Z. Wang, and Y. Xue, "Towards Building a Depression Diagnosis System Using Social Media Data," in Proceedings of the 2021 IEEE International Conference on Data Mining (ICDM), pp. 1238–1243.

J. D. Power and T. Mitchell, "Real-Time Prediction of Mental Health Disorders from Behavioral Data," IEEE Transactions on Computational Intelligence and AI in Games, vol. 13, no. 1, pp. 10–21, 2021.

L. Wang et al., "Sentiment Analysis of Depression Data Using Deep Learning Models," IEEE Transactions on Computational Social Systems, vol. 8, no. 4, pp. 857–865, Aug. 2021.

M. Shatte, D. Hutchinson, and P. Teague, "Machine Learning in Mental Health: A Scoping Review of Methods and Applications," Psychological Medicine, vol. 49, no. 9, pp. 1426–1448, 2019.

M. Amin and M. Rasheed, "Deep Learning for Mental Health Prediction: A Systematic Review," ACM Transactions on Computing for Healthcare, vol. 2, no. 4, pp. 1–21, 2021.

N. Milne, J. Moyle, S. Kellett, and T. W. Rogerson, "Application of Natural Language Processing in Psychiatric Text Analysis," Frontiers in Psychiatry, vol. 12, p. 580877, 2021.

P. Y. Chen, M. Bowring, and D. M. Fatima, "Deep Transfer Learning for Suicide Detection Using Text," in Proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2020, pp. 478–485.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter," arXiv preprint arXiv:1910.01108, 2019.

Y. Zhang, Y. Ma, and H. Liu, "Sentiment-Based Analysis for Mental Health Applications Using Transformer Models," IEEE Access, vol. 10, pp. 15468–15479, 2022.

Z. Huang et al., "Applications of Natural Language Processing in Clinical Psychiatry Research: A Systematic Review," Journal of Medical Internet Research, vol. 23, no. 10, p. e28465, 2021.