Detection of Mental Health Status of a Person using Text Data

Porika Rahul Naik *Indian Institute of Technology Bhubaneswar, School of Electrical Sciences, 22cs01068@iitbbs.ac.in*

Prof. Nitya Tiwari*Professor, School of Electrical Sciences, Indian Institute of Technology Bhubaneswar, India*

Abstract

Mental health disorders such as depression, anxiety, stress, and suicidal tendencies are increasingly being expressed through online platforms. Automated detection of mental health status from text data can assist in early intervention and clinical decision support. In this work, we present two deep learning-based approaches for multi-class mental health classification: (i) a **BiLSTM model with pretrained Word2Vec embeddings**, and (ii) a **transformer-based DistilBERT model**. Using a publicly available dataset from Kaggle containing labeled mental health-related text, we trained and evaluated both models. Experimental results demonstrate that the DistilBERT-based model achieved a classification accuracy of **G0.86%**, outperforming the BiLSTM + Word2Vec model. These results suggest that transformer-based language models capture contextual semantics more effectively, making them suitable for mental health text classification.

Keywords

Mental Health Detection, Natural Language Processing (NLP), Deep Learning, BiLSTM, Word2Vec, DistilBERT, Text Classification, Sentiment Analysis

1. Introduction

Mental health disorders represent one of the most pressing global health challenges, affecting millions of individuals worldwide. With the increasing prevalence of online forums, social media, and support communities, individuals often express their emotional states through textual posts. This provides an opportunity for **automatic detection of mental health conditions** using **Natural Language Processing (NLP)** techniques.

Traditional machine learning models relied on handcrafted features, but recent advancements in **deep learning** and **transformer architectures** have enabled more accurate detection by capturing semantic and contextual information in text.In this study, we investigate two models for multi-class mental health classification:

1. **BiLSTM with pretrained Word2Vec embeddings**, which leverages sequential context and distributed word representations.
2. **DistilBERT**, a lightweight transformer model capable of contextual understanding of text.

2. Literature Review

The detection of mental health conditions from text has attracted significant attention as individuals increasingly express emotions and psychological states on digital platforms. Researchers have applied Natural Language Processing (NLP) and deep learning techniques to identify cues related to depression, anxiety, stress, and suicidal tendencies.

Early research on detecting mental health conditions from text primarily relied on traditional machine learning algorithms. Models such as Support Vector Machines (SVM), Naïve Bayes, and Logistic Regression were widely adopted due to their simplicity and effectiveness. These approaches typically used handcrafted features like TF-IDF, bag-of-words, and sentiment lexicons to represent textual data. While they provided strong baselines and achieved reasonable accuracy, they struggled to capture deeper semantic meaning and contextual relationships between words. Such limitations restricted their ability to identify subtle psychological cues embedded in language. In particular, long-range dependencies and nuanced expressions often remained undetected. Consequently, these models lacked the sophistication needed for complex mental health text analysis. This created the need for more advanced approaches capable of learning richer linguistic representations.

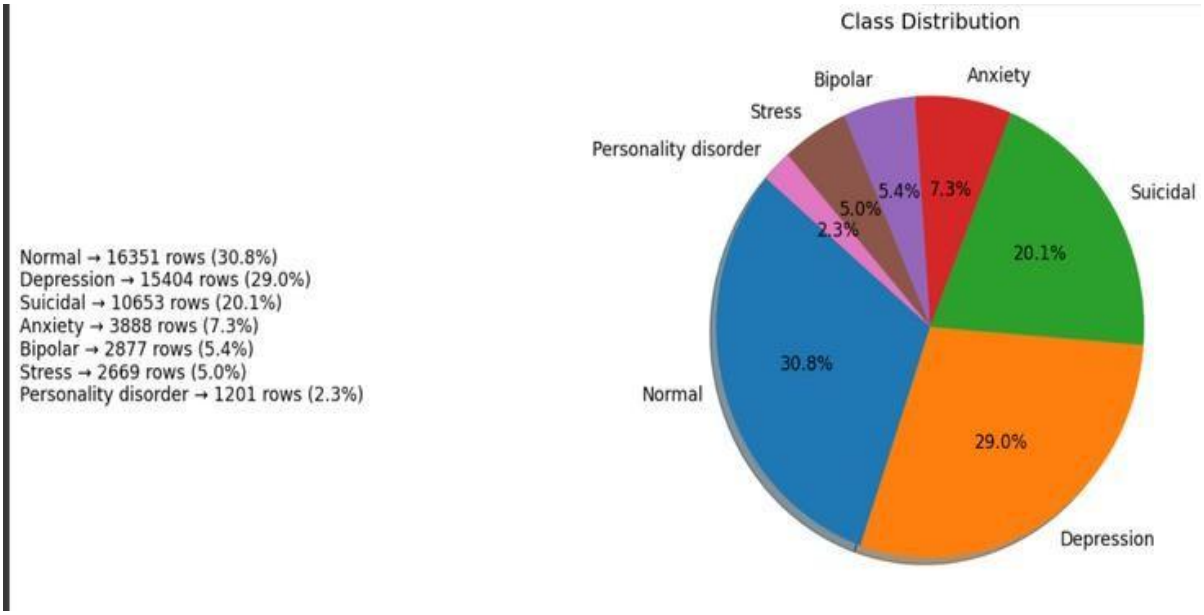
To overcome these drawbacks, word embeddings such as Word2Vec and GloVe were introduced, enabling the representation of semantic similarities. Combined with Recurrent Neural Networks (RNNs) and BiLSTM models, these approaches improved sequence modeling and demonstrated success in detecting mental health indicators from user-generated content.

In recent years, transformer-based architectures, particularly BERT and DistilBERT, have redefined text classification tasks. By employing self-attention mechanisms, they capture long-range dependencies and nuanced contexts more effectively. Literature consistently highlights their superior performance over earlier models, establishing transformers as the most promising approach for mental health text classification and early intervention.

3. Methodology

3.1 Dataset

We used the **Mental Health Sentiment Analysis dataset** from Kaggle [1], which contains labeled text data corresponding to different mental health conditions such as:



Class Distribution of Mental Health Dataset

*The chart illustrates the distribution of text samples across different mental health categories. The majority of data falls under the **Normal** class (16,351 samples, 30.8%), followed closely by **Depression** (15,404 samples, 29.0%). A significant portion is also labeled as **Suicidal** (10,653 samples, 20.1%). Smaller but important categories include **Anxiety** (3,888 samples, 7.3%), **Bipolar** (2,877 samples, 5.4%), **Stress** (2,669 samples, 5.0%), and **Personality disorder** (1,201 samples, 2.3%).*

Overall, the dataset is somewhat imbalanced, with a large proportion of samples concentrated in the Normal, Depression, and Suicidal classes. This imbalance could influence the performance of classification models, making it important to consider techniques like **resampling**, **class weighting**, or **data augmentation** to ensure that minority classes are adequately represented during model training. The diversity of categories, however, provides a comprehensive basis for developing models capable of detecting a wide range of mental health conditions from textual data.

3.2 Data Preprocessing

To prepare the dataset for modeling, the following preprocessing steps were applied:

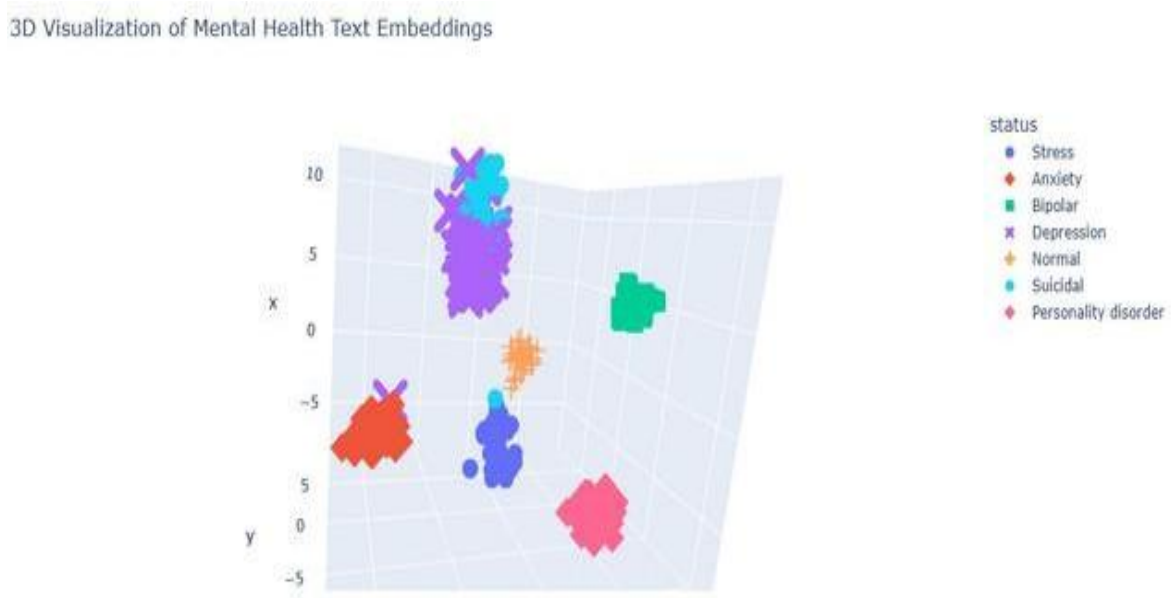
1. **Text Cleaning:** Removal of punctuation, URLs, hashtags, unwanted numbers, and special characters to retain only meaningful textual content.
2. **Spelling Correction:** Common spelling mistakes and informal abbreviations were corrected using a spell-checker and context-aware substitution.
3. **Tokenization and Normalization:** Text was converted to lowercase and tokenized into words/subwords depending on the model.
   - For **BiLSTM + Word2Vec**, standard word tokenization was used.
   - For **DistilBERT**, the pretrained WordPiece tokenizer was applied.
1. **Data Augmentation:** To enhance generalization and reduce overfitting, augmentation techniques were used:
   - **Back Translation:** Translating text to another language (e.g., English → French → English) and back to generate paraphrases.
   - **Paraphrasing:** Using pretrained NLP models to generate alternative formulations of the same sentence.
   - **Synonym Replacement:**
1. **Data Balancing:** The dataset was imbalanced, with classes like *depression* and *anxiety* dominating over *bipolar* or *personality disorder*. To overcome this, oversampling, augmentation (paraphrasing, back-translation), and limited undersampling were applied, ensuring a balanced class distribution for fairer model training.
2. **Padding/Truncation:** All sequences were normalized to a fixed maximum length of 128 tokens.
3. **Label Encoding:** Categorical labels were converted into numerical format and one-hot encoded for classification.

3.3 Word Embedding Layer

- **BiLSTM + Word2Vec:**The embedding layer was initialized with **pretrained Word2Vec vectors** (300 dimensions, trained on the Google News corpus). Each word token was mapped to its dense vector representation. Unknown or out-of-vocabulary words were assigned a special <UNK> token. These embeddings were **fine-tuned during training**, allowing the model to adapt them to the mental health domain.
- Word2Vec → **static embeddings** (same vector for a word regardless of context).
- **DistilBERT:**Unlike Word2Vec, DistilBERT uses **contextualized embeddings** generated by its transformer encoder. Each token (subword from WordPiece tokenization) is mapped to a contextual embedding that captures semantic

meaning based on surrounding words. The [CLS] token embedding from the final hidden layer was extracted and fed into the classification layer.

- DistilBERT → **contextual embeddings** (different vector depending on context).



3D Visualization of Mental Health Text Embeddings

The plot presents a three–dimensional projection of text embeddings derived from mental health statements. Each color and marker corresponds to a specific class, including stress, anxiety, bipolar disorder, depression, normal, suicidal, and personality disorder. Distinct clusters are visible, showing that sentences belonging to the same category share semantic similarities and are grouped closely together. At the same time, separation across clusters reflects meaningful differences among mental health conditions. This clustering pattern highlights the ability of the embedding model to capture contextual and emotional nuances in the text, making it suitable for classification and further analysis.

3.4 Model 1 : BiLSTM + Word2Vec

We designed a deep neural architecture that integrates sequential modeling, residual learning, and hierarchical attention to improve text classification performance.

1. **Input Representation**The model accepts tokenized text sequences and maps them into dense semantic vectors using a **300-dimensional Word2Vec embedding layer**. To enhance robustness against overfitting and noisy inputs, the embeddings are regularized with **SpatialDropout1D (rate = 0.3)** and **Gaussian noise injection (σ = 0.1)**. An additional *Augment Noise and Scale* step is applied to further improve generalization.

2. **Sequential Encoding with BiLSTM Layers**The processed embeddings are passed through **two stacked Bidirectional LSTM (BiLSTM) layers**, each containing 64 hidden units. These recurrent layers capture both past and future contextual dependencies within the sequence. To prevent overfitting and co-adaptation, a **Dropout layer (rate = 0.4)** is applied between the BiLSTM layers.

3. **Residual Learning**To ensure efficient gradient flow and preserve semantic information captured in earlier layers, a **residual connection** is introduced between the outputs of the first and second BiLSTM layers. This residual addition allows the model to leverage both low-level and high-level contextual representations.

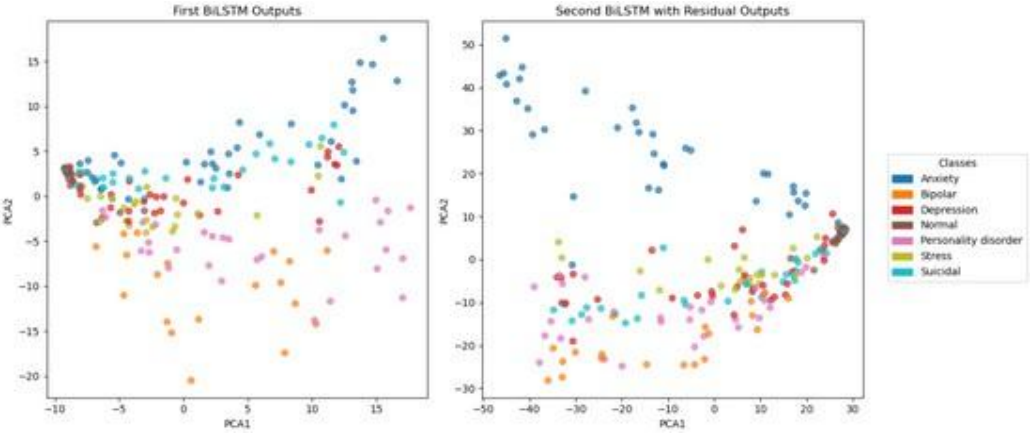


Figure 3.1

4. **Hierarchical Attention Mechanisms**A dual-level attention framework is employed to refine the extracted features:

- **Word-level attention** is applied to the output of the second BiLSTM, emphasizing informative tokens within a sentence.
- **Sentence-level attention** is applied to the first BiLSTM output, highlighting contextual cues across the sentence representation.

The two attention signals are then combined through **element-wise addition**, producing a unified representation that balances fine-grained token importance with broader sentence-level context.

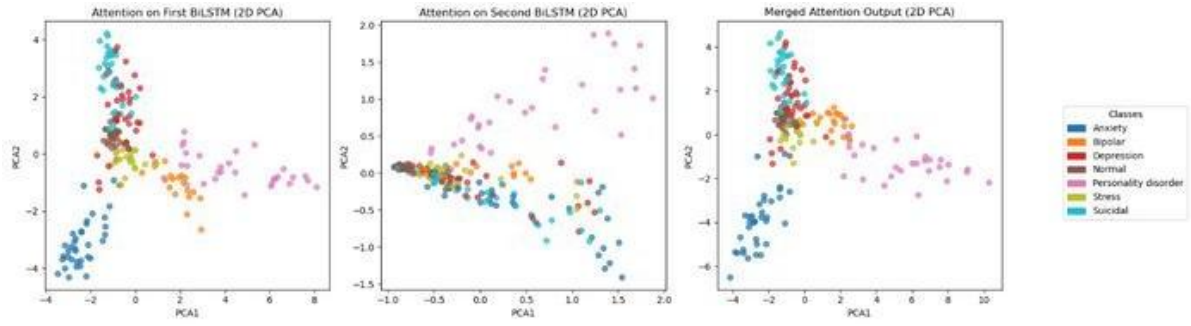


Figure 3.2

5. Normalization and Classification Layer The aggregated attention output is normalized using **Layer Normalization**, which stabilizes training and accelerates convergence. This is followed by a **fully connected dense layer** with 128 units, **ReLU activation**, and **L2 regularization (1e-5)**. A **Dropout layer (rate = 0.4)** is applied for additional regularization. Finally, classification is performed using a **Softmax output layer**, yielding probability distributions over the **7 target classes**.

Model Architecture Diagram :

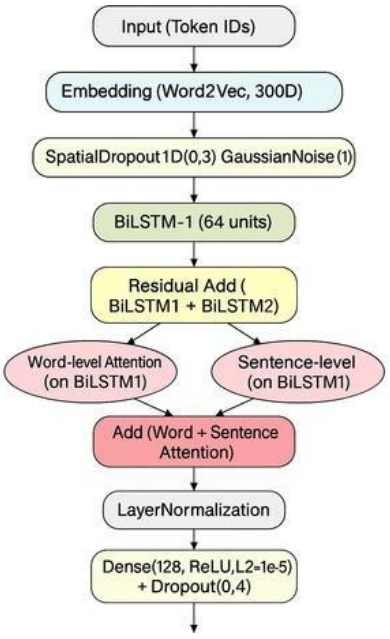


Figure 3.3 : BiLSTM + Word2Vec architecture with residual connection and hierarchical attention

Performance Summary

The proposed dual BiLSTM model with residual connections and dual-level attention achieved **8G.0G% training accuracy** and **85.64% validation accuracy**, with validation loss reduced to **0.61**. The small gap (~3.5%) between training and validation accuracy indicates good generalization.

Figure 3.4

Accuracy: 86%

- Best: Personality disorder (F1 = 0.96), Bipolar (0.92), Anxiety (0.91), Normal (0.90)
- Moderate: Stress (0.89)
- Lower: Depression (0.68), Suicidal (0.74)

Model performs well overall but struggles with Depression C Suicidal due to overlapping cues.

Strengths

- High accuracy with stable convergence.
- Residual + dual attention effectively capture semantic dependencies.
- Strong regularization reduces overfitting.

Limitations

- Computationally expensive (large training time, ~3.3M parameters).
- Validation accuracy plateaus after ~8 epochs.
- Complex architecture reduces interpretability and hinders deployment on low-resource systems.

3.5 Model 2: DistilBERT

1. Text Representation

For preprocessing, the model relies on the **DistilBERT tokenizer**, which uses WordPiece subword tokenization to handle rare and complex words effectively. Each sentence is truncated or padded to a maximum length of 128 tokens, and the tokenizer generates two key outputs: input_ids and attention_mask. These standardized inputs ensure that all text samples are represented in a format suitable for the model.

2. Model Architecture

The model architecture is based on **TFDistilBertForSequenceClassification**, a TensorFlow implementation of DistilBERT designed specifically for classification tasks. Pretrained weights from HuggingFace's **distilbert-base-uncased** are used to initialize the model, allowing it to benefit from transfer learning. On top of the base encoder, a classification head is added, with the number of output units set equal to the number of mental health categories in the dataset. The model produces logits that represent the raw scores for each class

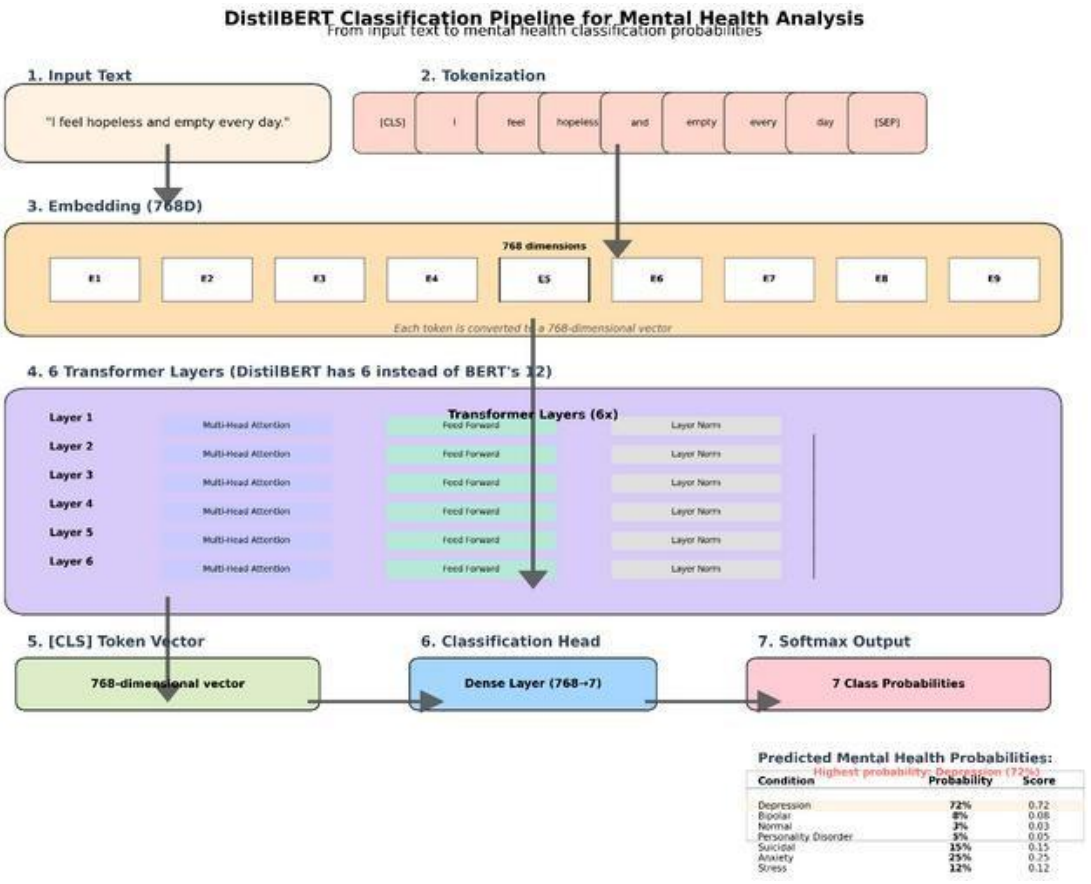


Figure 3.5

Training Setup

The model is trained using **Categorical Crossentropy loss** with logits, optimized using the **AdamW optimizer**. The optimizer is configured with a learning rate of 3e-5, a weight decay of 0.01, and a warm-up schedule covering 10% of the total training steps. Training is carried out for **3 epochs** with a **batch size of 16**, and the data pipeline is managed using tf.data.Dataset for efficiency. Model performance is evaluated on the validation set using **accuracy** as the metric.

Performance Summary

The DistilBERT-based mental health classification model was trained for 3 epochs and achieved **G4.7% training accuracy** and **G0.8% validation accuracy**. Validation loss decreased steadily, showing good generalization with minimal overfitting. The model can reliably classify mental health statements, as seen in the example where *"I feel hopeless and empty every day"* was predicted as **Depression** with **64% confidence**.

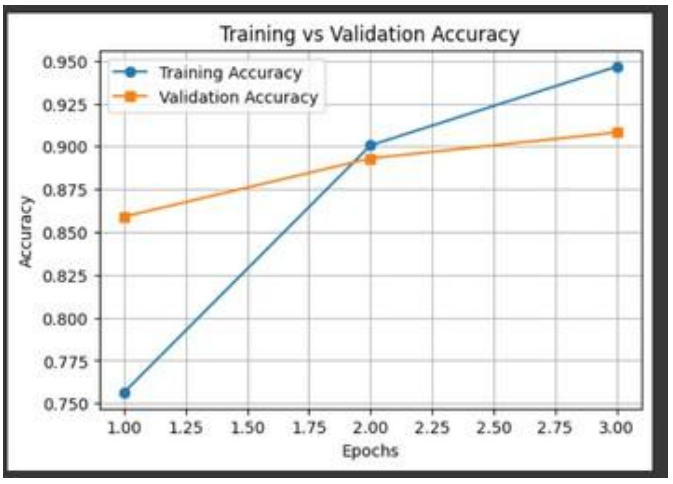


Figure -3.6

The graph in Figure 3.6 illustrates the trend of training and validation accuracy across three epochs. During the first epoch, both accuracies rise sharply, indicating that the model quickly learns fundamental language patterns relevant to mental health detection. In the subsequent epochs, the slope of improvement becomes more gradual, suggesting that the model shifts from rapid initial learning to fine-tuning more subtle contextual features. By the third epoch, validation accuracy surpasses 90%, closely following the training accuracy curve.

The small and consistent gap between the two curves demonstrates stable training with minimal overfitting, confirming the robustness and generalization capability of the DistilBERT model .

4. Results

Metric	DistilBERT Model	BiLSTM + Word2Vec Model
Accuracy	**90.82%**	85.00%
Best Performance Classes	Anxiety (0.96), Bipolar (0.97), Personality disorder (0.98)	Personality disorder (0.93), Normal (0.91), Anxiety (0.90)
Weakest Performance Classes	Depression (0.74), Suicidal (0.80)	Depression (0.68), Suicidal (0.76)
Precision-Recall Balance	More balanced across most classes	Significant precision-recall gaps in some classes
Training Efficiency	Faster convergence (3 epochs)	Slower convergence (11 epochs)
Architecture	Transformer-based (DistilBERT)	BiLSTM with Word2Vec embeddings
Required Resources	Higher memory and computation	Lower memory and computation

5. Conclusion

This study compared two deep learning models for detecting mental health status from text data: BiLSTM with pretrained Word2Vec embeddings and DistilBERT. Experimental results demonstrated that DistilBERT achieved superior performance, reaching **G0.8% accuracy**, compared to BiLSTM's ~86%.

These findings highlight the potential of transformer-based architectures in mental health detection tasks. Future work will focus on:

- Incorporating multimodal data (text + speech + physiological signals).
- Exploring larger pre-trained models like **RoBERTa** or **ClinicalBERT**.

References

[1] Kaggle dataset: *Mental Health Sentiment Analysis* – /kaggle/input/sentiment-analysis-for-mental-health/Combined Data.csv

https://www.kaggle.com/datasets/annastasy/mental-health-sentiment-analysis-nlp-ml[2] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9*(8), 1735-1780.[3] Devlin, J., et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT*.

[4] previous model link--
https://drive.google.com/file/d/1t5_Q1BDS5LX8LPDgiP2NBUsigPsFxZi_/view?usp=sharing