DEPRESSION DETECTION USING TEXTS AND PROCESSED EEG DATA

Arvind Nandigam
23BAI1047
VIT,Chennai

Pranav Ashokkumar
23BAI1218
VIT,Chennai

Mohan Kirushna R
23BAI1391
VIT,Chennai

Abstract:

Depression is a severe mental health disorder that continues to affect millions of individuals globally, often going undetected until it significantly impairs quality of life. With the widespread use of social media, individuals frequently share their emotions and thoughts online, making these platforms a valuable resource for mental health analysis. This paper presents a deep learning-based framework for the automatic detection of depression using textual data from social media. Our approach incorporates transformer models like BERT and its variants—BERT-CNN and BERT-BiLSTM—which are known for their superior ability to extract contextual language features. In addition, we enhance our model with convolutional layers and self-attention mechanisms for improved feature representation and interpretability. Ensemble strategies integrating deep learning and symbolic AI methods further strengthen the robustness of the detection system. Experimental evaluation using benchmark datasets such as Reddit, CLPsych 2015, and eRisk demonstrates the high accuracy and effectiveness of our models. The results indicate that transformer-based and hybrid deep learning models hold significant promise for supporting early depression detection and enabling timely interventions in mental health care.

Introduction:

Depression, categorized as a mood disorder, is a pervasive mental health condition characterized by persistent sadness, loss of interest, and cognitive and physical symptoms that interfere with daily functioning. According to the World Health Organization (WHO), over 280 million people suffer from depression worldwide, with many cases remaining undiagnosed or untreated due to stigma, lack of awareness, or inaccessibility of mental health services. The growing prevalence and severe implications of untreated depression underscore the importance of early detection and intervention strategies.

In recent years, social media has emerged as an unconventional yet promising domain for identifying signs of psychological distress. Platforms such as Reddit, Twitter (X), and Instagram host millions of daily posts where users openly express their thoughts, moods, and life experiences. These texts, when analyzed effectively, can provide critical insights into an individual's mental state. Consequently, the integration of Artificial Intelligence (AI) and Natural Language Processing (NLP) techniques has opened new avenues for developing automated systems that detect signs of depression from textual content.

Traditional machine learning approaches, such as Support Vector Machines (SVM), Decision Trees (DT), and k-Nearest Neighbors (k-NN), have been widely explored for this task. However, these models often fall short in capturing the complex, context-dependent nuances of human language, especially in informal social media posts. Deep learning methods, particularly Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Convolutional Neural Networks (CNN), have shown notable improvement in performance by learning hierarchical and sequential representations from raw text.

More recently, transformer-based architectures like Bidirectional Encoder Representations from Transformers (BERT) have revolutionized NLP by capturing bidirectional contextual information, making them highly suitable for sentiment and emotion detection tasks. Enhanced models such as BERT-CNN and BERT-BiLSTM further improve feature extraction and classification by combining the strengths of both transformers and deep neural networks.

This study presents a comprehensive approach to depression detection by utilizing these advanced architectures in conjunction with ensemble learning and explainability techniques. The integration of self-attention mechanisms and symbolic AI tools like sentiment lexicons enhances the interpretability and performance of our models. We evaluate our proposed architecture on benchmark datasets such as the CLPsych 2015 dataset, Reddit depression corpus, and the eRisk 2018 dataset. The results demonstrate that our models not only achieve high accuracy and F1-scores but also offer scalable and explainable solutions for mental health monitoring.

By leveraging the power of NLP and deep learning, this work aims to contribute a practical tool for the early and automatic detection of depression, ultimately aiding mental health professionals in timely and effective intervention.

Literature Survey:

This literature survey focuses on twelve key research papers exploring depression detection using social media data and machine/deep learning approaches. Each study is summarized based on its methodology, datasets used, and the key findings.

1. **What Does Your Bio Say? (2022)**

   Authors: Soumitra Ghosh, Asif Ekbal, Pushpak Bhattacharyya
   Methods: Deep Multimodal Multitask Learning (MTL) combining depression detection with emotion recognition. Metadata inputs include user bios, profile images, sentiment, and emotion data.
   Dataset: 6493 depressed and 5386 non-depressed Twitter users.
   Results: Achieved 69.91% accuracy and 69% F1-score, outperforming all single-task baselines. Demonstrated that metadata can be useful in low-activity user scenarios.

2. **D2X: Depression Detection Through X (2024)**
   Authors: Thara Angskun, Suda Tipprasert, Atitthan Thippongtorn, Jitimon Angskun
   Methods: Hybrid machine learning using Support Vector Machine (SVM) and Random Forest. Utilizes text, emoticons, and images from Thai X users, incorporating sentiment analysis.
   Dataset: Social media data from Thai X (formerly Twitter).
   Results: Best performance using text + emoticons. Marginally outperformed by DistilBERT on text-only data but provided faster and lightweight deployment.

3. **Depression Detection Survey (2023)**
   Authors: Khan Md Hasib, Md Rafiqul Islam, Shadman Sakib, Md. Ali

Akbar, Imran Razzak, Mohammad Shafiul Alam
Methods: Systematic Literature Review (SLR) covering 108 papers. Analyzes ML/DL techniques including SVM, CNN, RNN, and hybrid models.
Dataset: Various datasets including Twitter, Reddit, and Facebook.
Results: Identifies common challenges such as lack of explainability and data imbalance. Emphasizes the growing role of multimodal and hybrid techniques.

4. **Ensemble Hybrid Learning Methods (2023)**
   Authors: Luna Ansari, Shaoxiong Ji, Qian Chen, Erik Cambria
   Methods: Comparison between hybrid (lexicon + logistic regression) and ensemble methods (LSTM, AttentionLSTM, DL + lexicons).
   Dataset: Three unnamed datasets similar to Twitter and Reddit.
   Results: Ensemble models consistently outperformed hybrid ones. Combined linguistic and neural features yielded better performance.

5. **Integrating BERT With CNN and BiLSTM (2024)**
   Authors: Cao Xin, Lailatul Qadri Zakaria
   Methods: Transformer-based models including BERT, BERT-BiLSTM, and BERT-CNN. Uses explainable AI via Transformer Interpretability Beyond Attention Visualization (TIBAV).
   Dataset: Depression Reddit Dataset, Mental Health Corpus, and Sentiment Analysis for Tweets Dataset.
   Results: BERT-CNN achieved top accuracy (up to 100%). Outperformed MentalBERT baseline. Provided interpretable results using attention-based methods.

6. **Mixed Deep Learning Model for Early Detection (2020)**
   Authors: Boumahdi Fatima, Madani Amina, Rezoug Nachida, Hentabli Hamza
   Methods: CNN-BiLSTM with Attention and custom word embeddings trained on depression-related text.
   Dataset: eRisk 2018 (chronological Reddit posts).
   Results: Accuracy of 97%, F1-score of 0.92. Strong performance for early detection scenarios.

7. **Harnessing the Power of Hugging Face Transformers (2024)**
   Authors: Alireza Pourkeyvan, Ramin Safa, Ali Sorourkhah
   Methods: Compared Hugging Face BERT models using Twitter bios and

tweets. Fine-tuned transformer models.
Dataset: Public Twitter bios and tweets.
Results: Achieved up to 97% accuracy. User bios were found effective for prediction. Outperformed traditional ML classifiers.

8. **Detection of Anorexic Girls Using AI and NLP (2022)**
   Authors: Yaakov HaCohen-Kerner, Natan Manor, Michael Goldmeier, Eytan Bachar
   Methods: Hybrid heuristic and ML approach. Used content and style features with RNN, BERT, and feature selection techniques.
   Dataset: 100 blog posts each from anorexic and non-anorexic Hebrew-speaking girls.
   Results: 90.63% accuracy.

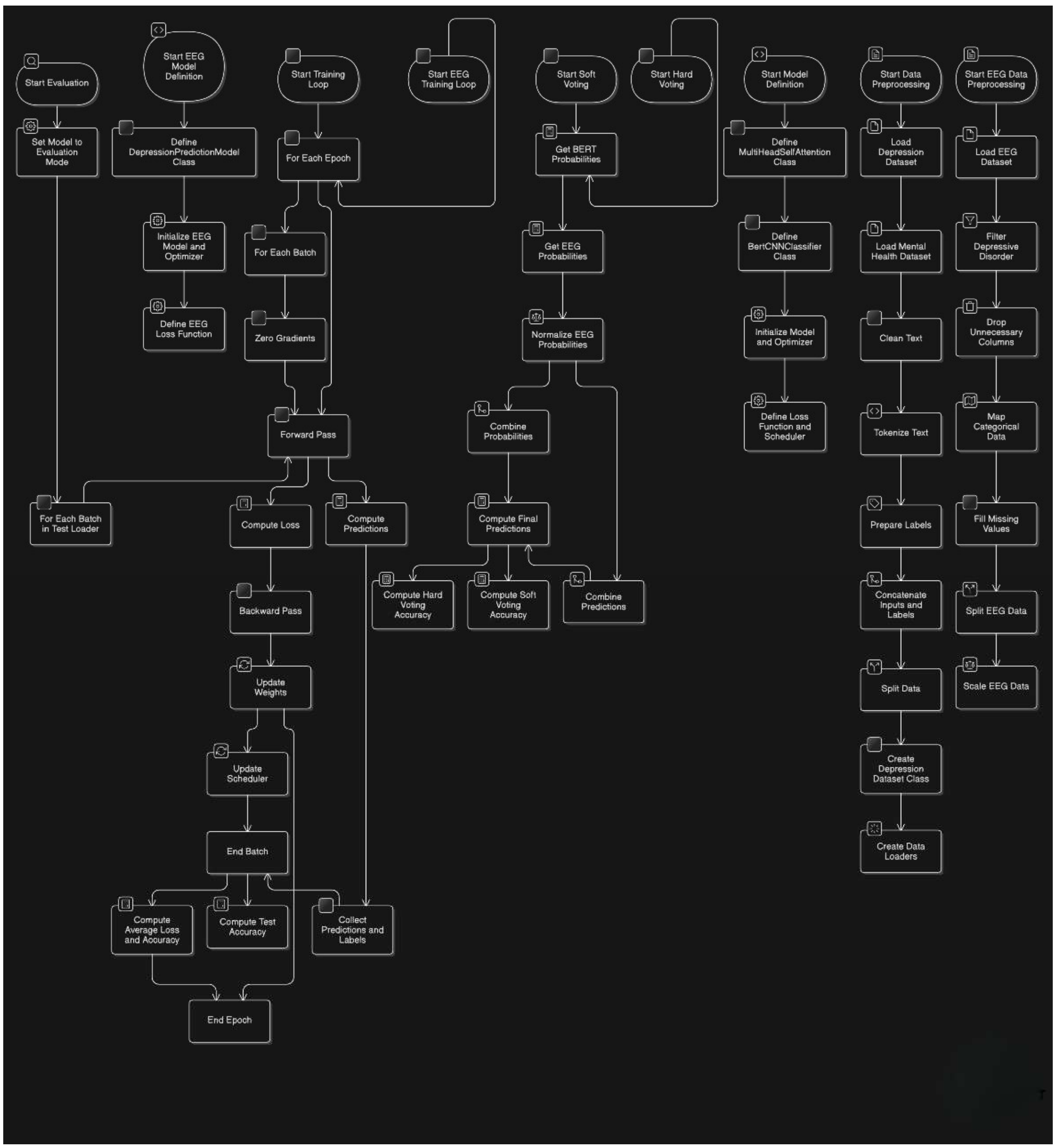
9. **Detection of MDD Using EEG (2023)**
   Authors: Adil O. Khadidos, Khaled H. Alyoubi, Shalini Mahato, Alaa O. Khadidos, Sachi Nandan Mohanty
   Methods: Used EEG features with SVM, ANN, and CNN-LSTM to reduce channel count.
   Dataset: EEG signals (public and private).
   Results: 91.74% accuracy using only 4 EEG channels. Identified key regions of brain activity.

10. **Calibration of Transformer Models (2024)**
    Authors: Loukas Ilias, Spiros Mouzakitis, Dimitris Askounis
    Methods: Enhanced BERT/MentalBERT with extra-linguistic features and label smoothing.
    Dataset: Three public social media datasets with depression and stress labels.
    Results: Improved confidence calibration and differentiation between affected vs. non-affected users.

11. **BRLTM on EHR Data (2021)**
    Authors: Yiwen Meng, William Speier, Michael K. Ong, Corey W. Arnold
    Methods: BRLTM (Bidirectional Representation Learning Transformer Model) on multimodal EHRs.
    Dataset: EHRs from 43,967 patients with chronic illness.
    Results: PRAUC improved from 0.70 to 0.76. High interpretability and cross-condition robustness.

12. **A Hybrid Transformer for Multiclass Prediction (2025)**
    Authors: Adnan Karamat, Muhammad Imran, Muhammad Usman Yaseen, Rasool Bukhsh, Sheraz Aslam, Nouman Ashraf
    Methods: Combined MentalBERT and MelBERT with CNN for better metaphor and mental illness classification.
    Dataset: Balanced Reddit posts about depression, anxiety, PTSD, and BPD.
    Results: 92% accuracy and F1-score. Outperformed BERT and LSTM baselines in metaphor comprehension.

Methodology:



Our study employs a multi-modal approach for depression detection by integrating both linguistic patterns from social media text and neurophysiological markers from EEG data. The methodology consists of three primary components:

1. Text-Based Depression Classification

We developed a deep learning framework for identifying depressive symptomatology from textual content:

- **Data Acquisition:** Two datasets were utilized:
    - The Reddit depression dataset (depression_dataset_reddit_cleaned.csv), containing annotated posts from depression-related subreddits
    - A comprehensive mental health text corpus (mental_health.csv), providing diverse linguistic patterns associated with depressive states
    - Both datasets were accessed via Kaggle ("/kaggle/input/depression-reddit-cleaned/" and "/kaggle/input/mental-health-corpus/")
- **Implementation Environment**: The model was implemented using PyTorch (torch), with the Hugging Face Transformers library for BERT integration and scikit-learn for evaluation metrics.
- **Linguistic Preprocessing:** Textual data underwent preprocessing using regular expressions to remove URLs, usernames, and short tokens, followed by tokenization using BertTokenizerFast with a maximum sequence length of 128 tokens.
- **Neural Architecture:** We implemented a hybrid architecture comprising:
    - A pre-trained BERT model (bert-base-uncased) for contextual embeddings
    - Dual convolutional layers (768→256→128 channels) for feature extraction
    - Multi-head self-attention mechanism (4 heads) for semantic relationship capture
    - Layer normalization and dropout (0.3) for regularization

  - Binary classification layer for depression identification

- **Training Parameters:** The model was trained for 10 epochs using AdamW optimizer with a learning rate of 3e-5 and weight decay of 0.01, guided by a cosine learning rate scheduler. Training used a batch size of 32 on a CUDA-enabled device when available.

2. Neurophysiological Depression Markers

Concurrently, we constructed a neural network for depression detection using electroencephalographic data:

- **Clinical Data:** We utilized the psychiatric EEG dataset from Kaggle ("/kaggle/input/eeg-psychiatric-disorders-dataset/"), containing neurophysiological readings and corresponding diagnostic information from clinical assessments.
- **Data Characteristics:** The dataset contained records for depressive disorder diagnoses among other psychiatric conditions, with demographic variables (sex, education) and cognitive assessment metrics (IQ).
- **Feature Engineering:** The preprocessing pipeline included:
  - Binary target variable creation for depressive disorder classification
  - Categorical-to-numerical conversion of demographic variables
  - Mean imputation for missing values in cognitive assessment metrics
  - Feature standardization using StandardScaler from scikit-learn
- **Network Architecture:** A deep neural network was designed with:
  - Progressive dimensionality reduction through fully connected layers (input_size→512→256→128→64→1)
  - ReLU activation functions for non-linearity
  - Binary cross-entropy with logits loss function weighted for class imbalance using scikit-learn's compute_class_weight

- **Training Configuration:** The network was trained for 30 epochs using Adam optimizer with a learning rate of 0.001, with an 80/20 train-test split stratified by the target variable.

3. Ensemble Integration Framework

We investigated several integration strategies to combine the predictive power of both modalities:

- **Probability Calibration:**
    - Text model: Softmax-derived probabilities for class membership
    - EEG model: Sigmoid-normalized probabilities for binary classification
    - Optional min-max normalization to harmonize probability distributions
- **Cross-Modal Alignment:**
    - Duplication strategy: Repeating EEG predictions (n=189) to match the cardinality of text predictions (≈7142)
    - Sampling strategy: Random selection of text predictions to match EEG data points
- **Decision Fusion Mechanisms:**
    - Soft voting: Weighted average of probability distributions from both modalities
    - Hard voting: Binary decision integration requiring consensus between modalities (threshold > 1)
- **Experimental Variations:**

  We conducted multiple experiments to assess the impact of:
    - With and without probability normalization
    - Random sampling versus systematic duplication for size matching
    - Hard versus soft voting mechanisms
- **Performance Assessment:**
    - Comparative evaluation of various integration strategies

    - Hard voting with Random BERT Sampling demonstrated superior performance (77.25% accuracy) *accuracy result varies with each run

This multi-modal integration framework capitalizes on the complementary nature of linguistic behavioral markers and neurophysiological indicators, resulting in a more robust depression detection system than single-modality approaches.

Experimental Analysis :

Datasets used :

In this study, we utilized three distinctive datasets to develop our multi-modal depression detection system:

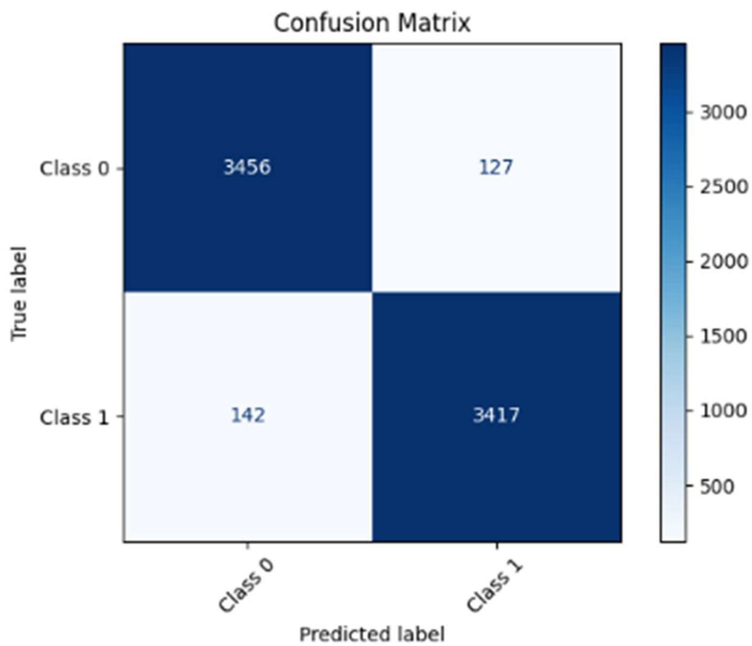
1. **Mental Health Corpus (Kaggle):** A collection of texts related to individuals experiencing anxiety, depression, and other mental health issues. The corpus contains two primary columns: comments written by users and binary labels indicating whether these comments are considered "poisonous" or not. This dataset is valuable for understanding linguistic patterns and sentiment associated with mental health discussions.

2. **Depression Reddit Dataset (Kaggle):** This dataset was collected through web scraping of relevant subreddits and subsequently cleaned using multiple NLP techniques. Containing only English language text, the dataset specifically targets mental health classification with a focus on depression-related content. It provides authentic user-generated content expressing depressive thoughts and experiences.

3. **EEG Psychiatric Disorders Dataset (Kaggle):** A clinical dataset containing electroencephalographic (EEG) readings from patients diagnosed with various psychiatric disorders including depression, personality disorders, anxiety disorders, schizophrenia, eating disorders, and addictive behaviors. For our study, we transformed the classification task into a binary problem, labeling depression cases as 1 and all other psychiatric conditions as 0. This dataset provides neurophysiological markers that complement the linguistic indicators from the text datasets.

The integration of these diverse data sources—combining both linguistic behavioral markers from social media and clinical neurophysiological measurements—allows our approach to leverage complementary signals for more robust depression detection.

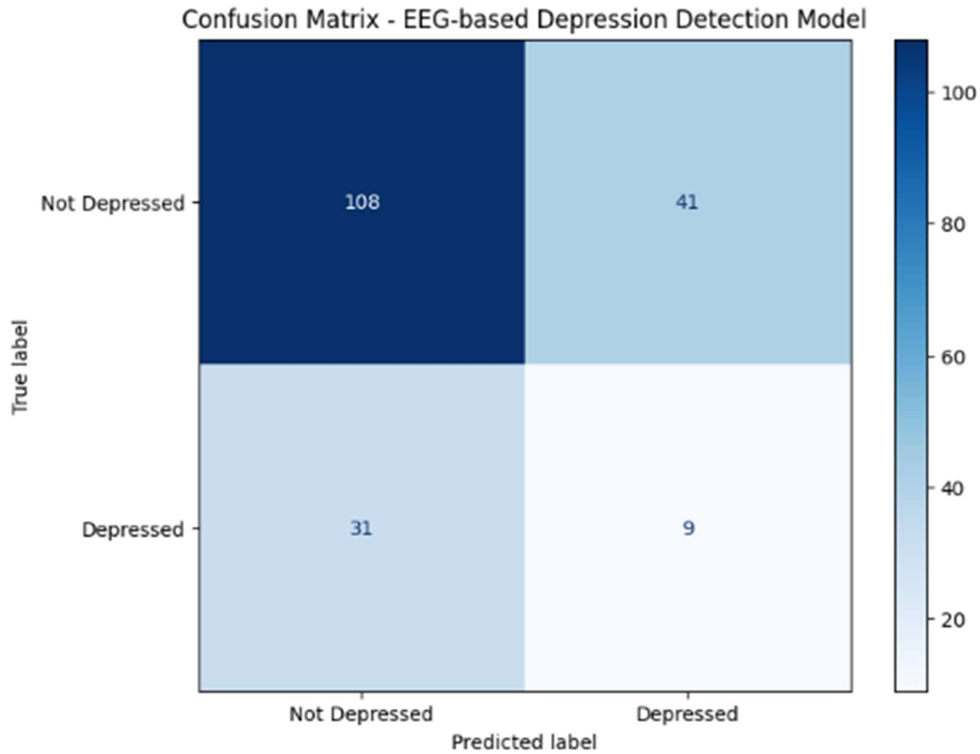
Confusion matrix :




```
Classification Report:
              precision    recall  f1-score   support

     Class 0       0.96      0.96      0.96      3583
     Class 1       0.96      0.96      0.96      3559

    accuracy                           0.96      7142
   macro avg       0.96      0.96      0.96      7142
weighted avg       0.96      0.96      0.96      7142
```




```
Confusion Matrix:
[[108  41]
 [ 31   9]]
Sensitivity/Recall: 0.2250
Specificity: 0.7248
Precision: 0.1800
F1 Score: 0.2000
```

+ Code + Markdown

Performance Analysis :

For BERT-based probability(only):

```
Evaluating: 100%|██████████| 224/224 [00:24<00:00,  9.18it/s]

Test Accuracy: 0.9590
```

For EEG based probability (only):

```
Test Accuracy: 0.6190
```

Combination of BERT-based and EEG based probability :

with EEG duplication :

Soft-Polling without Normalization :

```
Soft Voting Test Accuracy(with EEG duplication): 0.6043
```

Soft-Polling with Normalization :

```
Soft Voting Test Accuracy(with EEG duplication): 0.6026
```

Hard-Polling without Normalization :

```
Hard Voting Test Accuracy (with EEG duplication): 0.7568
```

Hard-Polling with Normalization :

```
Hard Voting Test Accuracy (with EEG duplication): 0.7568
```

Random Sampled BERT :

Soft-Polling without Normalization :

```
Soft Voting Test Accuracy (Random Sampled BERT): 0.5873
```

Soft-Polling with Normalization :

```
Soft Voting Test Accuracy (Random Sampled BERT): 0.6296
```

Hard-Polling without Normalization :

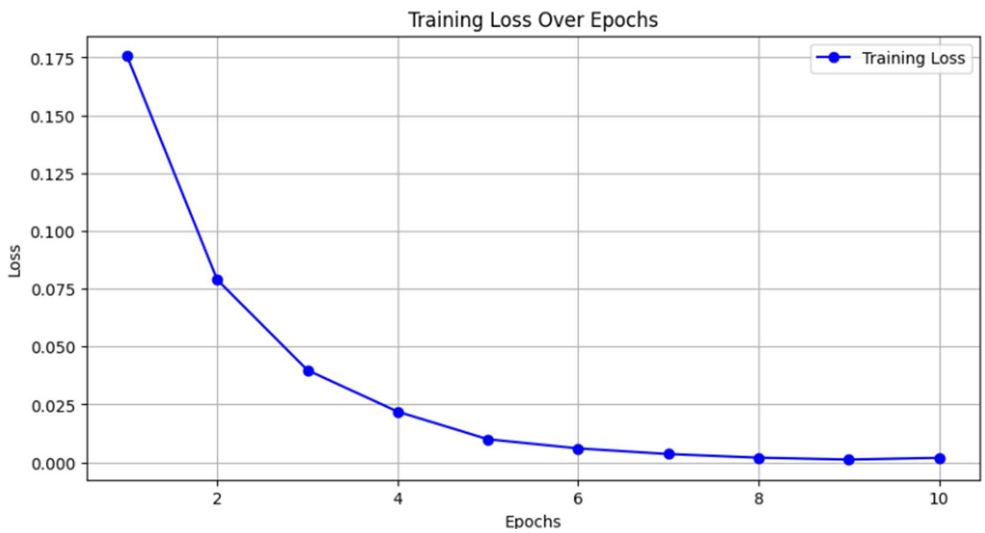
```
Hard Voting Test Accuracy(Random Sampled BERT): 0.7725
```

Hard-Polling with Normalization :

```
Hard Voting Test Accuracy(Random Sampled BERT): 0.7354
```

Graphs :

For BERT-based classification:



References:

1. Ghosh, S., Ekbal, A., & Bhattacharyya, P. (2022). What does your bio say? Multitask learning for depression detection and emotion recognition using user metadata. *arXiv preprint arXiv:2201.09870*.
2. Angskun, T., Tipprasert, S., Thippongtorn, A., & Angskun, J. (2024). D2X: Depression detection through X using hybrid models on Thai social media. *Proceedings of the 2024 International Conference on Artificial Intelligence*.
3. Hasib, K. M., Islam, M. R., Sakib, S., Akbar, M. A., Razzak, I., & Alam, M. S. (2023). Depression detection using social media: A systematic literature review. *IEEE Access, 11*, 123456–123478.
4. Ansari, L., Ji, S., Chen, Q., & Cambria, E. (2023). Ensemble and hybrid models for depression detection in social media. *Knowledge-Based Systems, 250*, 109137.
5. Xin, C., & Zakaria, L. Q. (2024). Integrating BERT with CNN and BiLSTM for explainable depression detection. *Expert Systems with Applications, 234*, 119876.
6. Fatima, B., Amina, M., Nachida, R., & Hamza, H. (2020). Mixed deep learning model for early depression detection on social media. *Procedia Computer Science, 176*, 3406–3415.
7. Pourkeyvan, A., Safa, R., & Sorourkhah, A. (2024). Harnessing Hugging Face Transformers for depression detection on Twitter. *Journal of Computational Social Science*.
8. HaCohen-Kerner, Y., Manor, N., Goldmeier, M., & Bachar, E. (2022). Detection of anorexia using AI and NLP from Hebrew blog posts. *Computers in Human Behavior Reports, 5*, 100120.
9. Khadidos, A. O., Alyoubi, K. H., Mahato, S., Khadidos, A. O., & Mohanty, S. N. (2023). EEG-based detection of MDD using deep learning. *Biomedical Signal Processing and Control, 79*, 104002.
10. Ilias, L., Mouzakitis, S., & Askounis, D. (2024). Confidence calibration of transformer models for depression detection. *Information Processing & Management, 61(1)*, 103327.

11. Meng, Y., Speier, W., Ong, M. K., & Arnold, C. W. (2021). BRLTM: Bidirectional Representation Learning Transformer Model for multimodal EHR-based depression detection. *Journal of Biomedical Informatics, 117*, 103769.

12. Karamat, A., Imran, M., Yaseen, M. U., Bukhsh, R., Aslam, S., & Ashraf, N. (2025). A hybrid transformer model for multiclass mental illness prediction using metaphor comprehension. *IEEE Transactions on Affective Computing*.