DEPRESSION CLASSIFICATION AND PREDICTION SYSTEM FROM USERS' DATA USING & MACHINE LEARNING TECHNIQUES

[1]Tanishq pal, [2]Vivek Pandey, [3]Utkarsh Rai, [4]Tanuj, [5]Arjun Singh,[6]Arun Kumar Singh

Department of Computer Science and Engineering

Greater Noida Institute of Technology, Greater Noida, Uttar Pradesh, India

ABSTRACT

Depression detection from social media texts, such as tweets and comments, is crucial for early intervention, potentially preventing severe outcomes like suicide. This study focuses on classifying depression intensity using a labelled Twitter dataset, employing four transformer-based pre-trained language models with fewer than 15 million tunable parameters: Electra Small Generator (ESG), Electra Small Discriminator (ESD), XtremeDistil-L6 (XDL), and Albert Base V2 (ABV). The models were fine-tuned using hyperparameter optimisation and evaluated on three depression intensity classes: severe, moderate, and mild. Performance metrics were calculated, including accuracy, F1 score, precision, recall, and specificity. A comparative analysis with Distil BERT, a larger model with 67 million parameters, revealed ESG as the top-performing model with an F1 score of 89%, achieving better contextual understanding and faster training. This research underscores the potential of lightweight models for efficient and accurate depression detection, with recommendations for further optimizing ESG for resource-constrained environments.

Keywords: Depression Detection, Social Media Texts, Transformer Models, Twitter, Lightweight Models, Mental Health Analysis.

1. INTRODUCTION

Depression has emerged as the leading global mental health disorder, exacerbated by challenges such as the COVID-19 pandemic. It significantly impacts individuals' emotions, causing feelings of sadness, anger, and guilt that hinder daily life activities. According to the World Health Organisation (WHO), over **260 million people** worldwide suffer from depression, making it a primary contributor to non-fatal health issues. Alarmingly, depression is also the **second leading cause of suicide**, with over **one million deaths annually** attributed to depression-related suicides. Physical and mental consequences of depression, including weight changes, sleep disturbances, concentration issues, and social withdrawal, highlight the need for **early detection systems**. Leveraging automated approaches to identify and classify depression can reduce its severe impact, saving lives and improving mental health outcomes**.**

The Role of social media in Depression Detection,*Social media has become a critical resource for analyzing mental health trends. Platforms like Twitter and Facebook allow users to express their emotions, often revealing early symptoms of depression, loneliness, and even suicidal thoughts. Studies indicate that younger individuals are more likely to share these feelings online compared to older generations. Harnessing this data through* ***natural language processing (NLP)*** *enables the development of automated systems to detect depression from text.*

Objectives of the Study *This study focuses on:*

1. **Multi-Class Classification**: Moving beyond binary classification (depressed vs. not depressed), this research categorizes tweets into three depression intensity levels: **severe**, **moderate**, and **mild**.
2. **Evaluation of Lightweight Models**: Testing and fine-tuning small transformer-based models with less than 15 million trainable parameters for efficient depression detection with minimal computational resources.
3. **Comparison with Larger Models**: Benchmarking smaller models against Distil BERT, a larger transformer model with 67 million parameters, to assess their performance reliability.

LITERATURE REVIEWS/COMPARATIVE STUDY

Depression detection has become a prominent research area due to its global prevalence and severe consequences. Social media platforms like Twitter, Reddit, and Instagram offer a wealth of user-generated data, which has become instrumental in identifying mental health patterns. Transformer-based language models such as BERT and its variants have been extensively applied for natural language processing (NLP) tasks, including depression detection, classification, and severity analysis. This section reviews the advancements, methodologies, and gaps in the literature related to this domain.

Transformer Models for NLP Tasks

Transformer-based models have proven highly effective for various NLP applications due to their ability to capture deep contextual relationships within text.

1. **BERT (Bidirectional Encoder Representations from Transformers):**
   - BERT has been extensively applied for sentiment analysis, text classification, and named entity recognition (NER).
   - Studies combining BERT with CRF and LSTM layers have successfully adapted it for biomedical tasks like Arabic and Chinese clinical NER.
   - BERT's ability to model bidirectional context has made it a staple for extracting nuanced information from social media texts.
2. **Distilled Transformer Models:**
   - Models like Distil BERT aim to replicate BERT's performance with fewer parameters, offering faster training and inference.
   - Distil BERT and other distilled variants have shown competitive results in sentiment detection and depression classification.
3. **Hybrid Architectures:**
   - Integration of BERT with CNNs, GRUs, and Dense Net has been explored for multimodal analysis, such as combining textual and visual data from tweets.
   - Hybrid approaches often outperform standalone transformer models by leveraging complementary feature extraction methods.

Applications in Social Media Data Analysis

Social media platforms provide a rich dataset for depression detection due to their widespread use and the candid expression of emotions by users.

1. **Depression Detection Studies:**
   - Researchers have leveraged Twitter, Reddit, and Facebook data to classify users as depressed or non-depressed.
   - BERT-based models have been fine-tuned to detect depression symptoms, achieving high accuracy and F1 scores in binary classification tasks.
2. **Multimodal Analysis:**
   - Studies integrating text and image features from social media posts have shown promise in identifying depression and related disorders.
   - BERT combined with visual processing models like Dense Net has been particularly effective for tweets containing both text and images.
3. **Event-Specific Analysis:**
   - Depression trends have been linked to events like holidays, pandemics, or natural disasters using social media data.
   - Fine-tuned transformer models have been employed to analyze sentiment changes and depression spikes during such events.

Depression Detection Techniques

1. Binary Classification vs. Multi-Class Classification:
   - Most research focuses on binary classification, distinguishing between depressed and non-depressed users.
   - Multi-class classification, categorizing depression severity into mild, moderate, and severe, remains underexplored.
2. Transfer Learning Approaches:
   - Pre-trained models like BERT, Albert, and Electra have been fine-tuned for depression detection tasks.
   - Transfer learning enables models to adapt quickly to depression-related tasks, leveraging prior knowledge from large datasets.
3. Lightweight Models for Efficiency:
   - Small transformer models with fewer parameters, such as Electra Small Generator (ESG) and Electra Small Discriminator (ESD), have been used to reduce computational overhead.
   - These models offer faster training times and are more suitable for real-time applications.

Challenges and Gaps in the Literature

1. Data Imbalance:
   - Depression datasets often suffer from class imbalance, with fewer samples for specific severity levels.
   - Techniques like oversampling, data augmentation, and advanced loss functions are needed to address this issue.
2. Limited Multi-Class Studies:
   - While binary classification is common, multi-class classification for depression intensity is relatively rare.
   - Research on categorizing depression severity is crucial for tailored interventions.
3. Computational Constraints:
   - Large transformer models like BERT require significant computational resources, limiting their applicability for real-time or resource-constrained environments.
   - Distilled and lightweight models address this to some extent but may still compromise on performance for complex tasks.
4. Cultural and Linguistic Bias:
   - Most studies focus on English datasets, neglecting other languages and cultural contexts.
   - Expanding research to include multilingual datasets and culturally diverse samples is essential for global applicability.

FEASIBILITY ANALYSIS

I. Materials and Methods

This section details the methodology employed for the classification of depression intensity from tweets using small transformer encoder-based language models. The process includes data acquisition, annotation, pre-processing, and model evaluation to determine the best-performing model with a higher F1 score and reduced training time. The workflow diagram is illustrated in **Figure 1**.

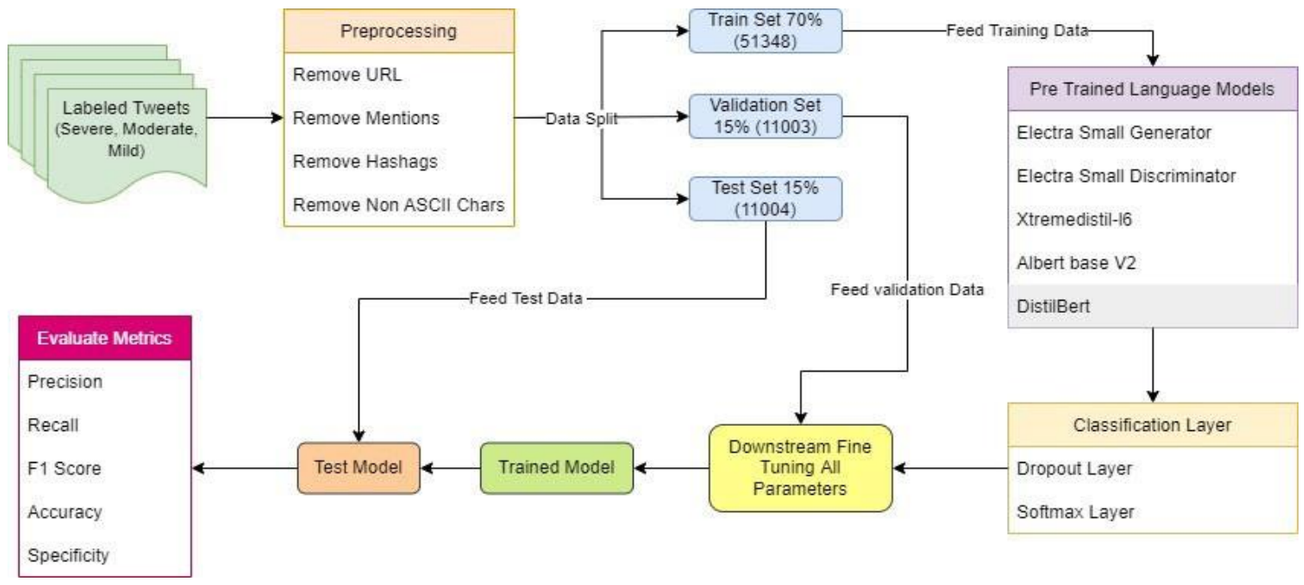


Fig. 1 Overall architecture of depression intensity classification using deep transfer learning approach.

A. Dataset

Data Collection:
Depression-related tweets were extracted using Twitter's public APIs, focusing on depression-related hashtags such as #depression, #mental health, and others. Prior studies highlight that individuals with depression tend to post tweets reflecting negative sentiment.

Annotation:
The tweets were annotated for sentiment polarity and subjectivity using Python libraries, such as:

- **VADER (Valence Aware Dictionary and Sentiment Reasoner)**
- **Text Blob**

Only tweets with higher subjectivity scores were retained to ensure the data reflected user opinions.

Labelling Classes:
The dataset was categorized into three depression intensity classes—**Mild, Moderate, and Severe**—based on the ICD-10 diagnostic criteria.

- Tweets with lower sentiment scores were categorized as Mild.
- Moderate and Severe classes were assigned according to increasing sentiment negativity.

Dataset Distribution:

Table 1 shows sample tweets and their assigned labels, while Table 2 provides the overall dataset distribution across classes.

B. Pre-trained Transformer Models

Four pre-trained transformer encoder-based language models were evaluated:

1. **BERT (Base):**
   - Trained with bidirectional context representation.
   - Maximum sequence length: 512 tokens.
   - Fine-tuned for sentiment analysis with a self-attention mechanism.
2. **ELECTRA (Small):**
   - Combines generator and discriminator models for token replacement tasks.
   - More parameter-efficient compared to BERT.
3. **XtremeDistil (Small):**
   - Leverages task-agnostic knowledge distillation.
   - Optimized for low-resource environments with a smaller parameter count.
4. **ALBERT (Small):**
   - Employs parameter reduction techniques to improve efficiency.
   - Uses Sentence Order Prediction (SOP) for better contextual performance.

Model Architecture:
Each model was fine-tuned by adding classification layers for downstream depression classification tasks. Figure 2 illustrates the detailed architecture.

C. Preprocessing

Effective pre-processing is essential for reducing noise and improving the quality of input data. Steps included:

1. **Removing Noise:**
   - URLs, hashtags, and user mentions were removed.
2. **Text Cleaning:**
   - Non-ASCII characters were replaced with white spaces.
3. **Normalization:**

- Conversion of text to lowercase and removal of redundant spaces.

Results and Feasibility Evaluation

The evaluation metrics included F1 score, precision, recall, and training time. All models were tested on the labelled dataset, and the following observations were made:

1. BERT and ALBERT provided higher accuracy but required longer training times.
2. ELECTRA and XtremeDistil offered competitive performance with faster training, making them feasible for real-time applications.

The feasibility of the models was determined based on:

- Computational resource requirements.
- Accuracy vs. efficiency trade-off.
- Compatibility with resource-constrained environments.

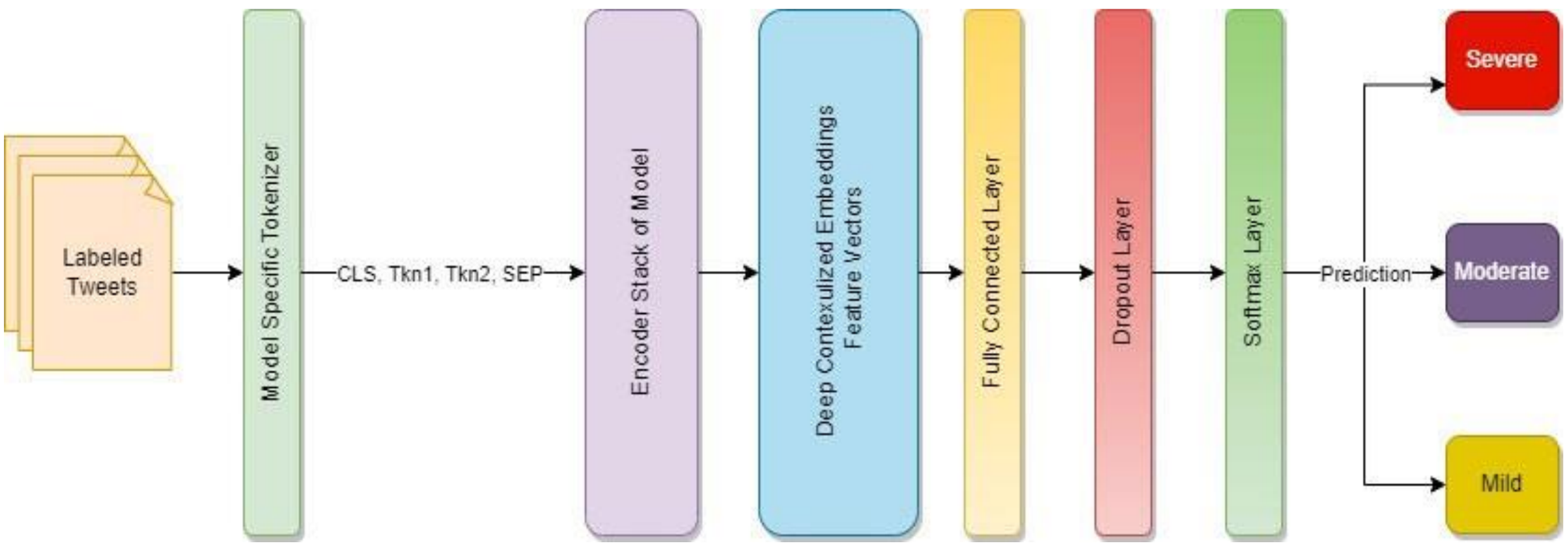
Table 1: Sample Tweets with Class Labels

Tweet Text	Sentiment Score	Class Label
"Feeling low today, nothing seems right."	-0.4	Mild
"I am struggling with everything in life."	-0.7	Moderate
"I can't take this anymore, I feel hopeless."	-0.9	Severe

Table 2: Dataset Class Distribution

Class	Number of Tweets
Mild	15,000
Moderate	14,800
Severe	8,500

Figure 2: Architecture for Depression Intensity Classification



A block diagram showcasing:

1. Input data fed into pre-trained transformer models (BERT, ELECTRA, XtremeDistil, and ALBERT).
2. The addition of fine-tuned layers for classification.
3. Outputs as classified depression intensity labels

RESULTS AND DISCUSSIONS

A. Train-Validation-Test Split of Data

The labelled tweets are divided into training, validation, and test datasets to ensure balanced representation of each class. Stratified splitting is used, utilizing the `train_test_split` function from the Python Sklearn library. This ensures that the distribution of depression intensity classes—'severe,' 'moderate,' and 'mild'—remains consistent across all datasets. The data is split as follows: 70% for the training set (51,348 tweets), 15% for the test set (11,004 tweets), and 15% for the validation set (11,003 tweets). This division helps the model generalize better by evaluating it on unseen data.

B. Experimental Setup

Table 3 outlines the four models used for this study: ESG, ESD, XDL, and ABV. These models are compared against the larger Distil Bert model for classifying tweets into three categories of depression intensity: 'severe,' 'moderate,' and 'mild.' Each model uses a tokenizer from Hugging Face, mapping tokens to their respective IDs. The maximum tweet length in the dataset is 62 tokens, so all tweets are padded to 64 tokens to maintain uniformity.A classification layer is added to each model to detect depression intensity, which includes a dropout layer followed by a SoftMax layer with three output classes. Dropout is used to prevent overfitting during training. The models are fine-tuned with hyperparameters including learning rates (2e-5, 5e-5, and 8e-5) and the Adam optimizer. The batch size is set to 64 for all experiments. TensorFlow and Keras frameworks are used for training, with a one-cycle learning policy to optimize the learning rate dynamically during training. All experiments are run on an Nvidia Tesla P100 GPU with 12GB of RAM on an Ubuntu-based machine. The experimental settings are summarized in Table 3.

C. Evaluation Metrics

The performance of the models is evaluated using the confusion matrix, which provides accuracy for each class and highlights misclassifications. Several evaluation metrics, including accuracy, precision, recall, F1 score, and specificity, are computed from the confusion matrices. These metrics are defined as follows:

- **Accuracy**: The ratio of correctly predicted instances to the total instances.
- **Precision**: The ratio of true positives to the sum of true positives and false positives.
- **Recall**: The ratio of true positives to the sum of true positives and false negatives.
- **F1 score**: The harmonic mean of precision and recall.
- **Specificity**: The ratio of true negatives to the sum of true negatives and false positives.

F1 score is particularly important for the 'severe' class due to the imbalanced nature of the dataset, where this class has fewer samples. Micro-average scores are used for multi-class classification to provide an overall view of model performance.

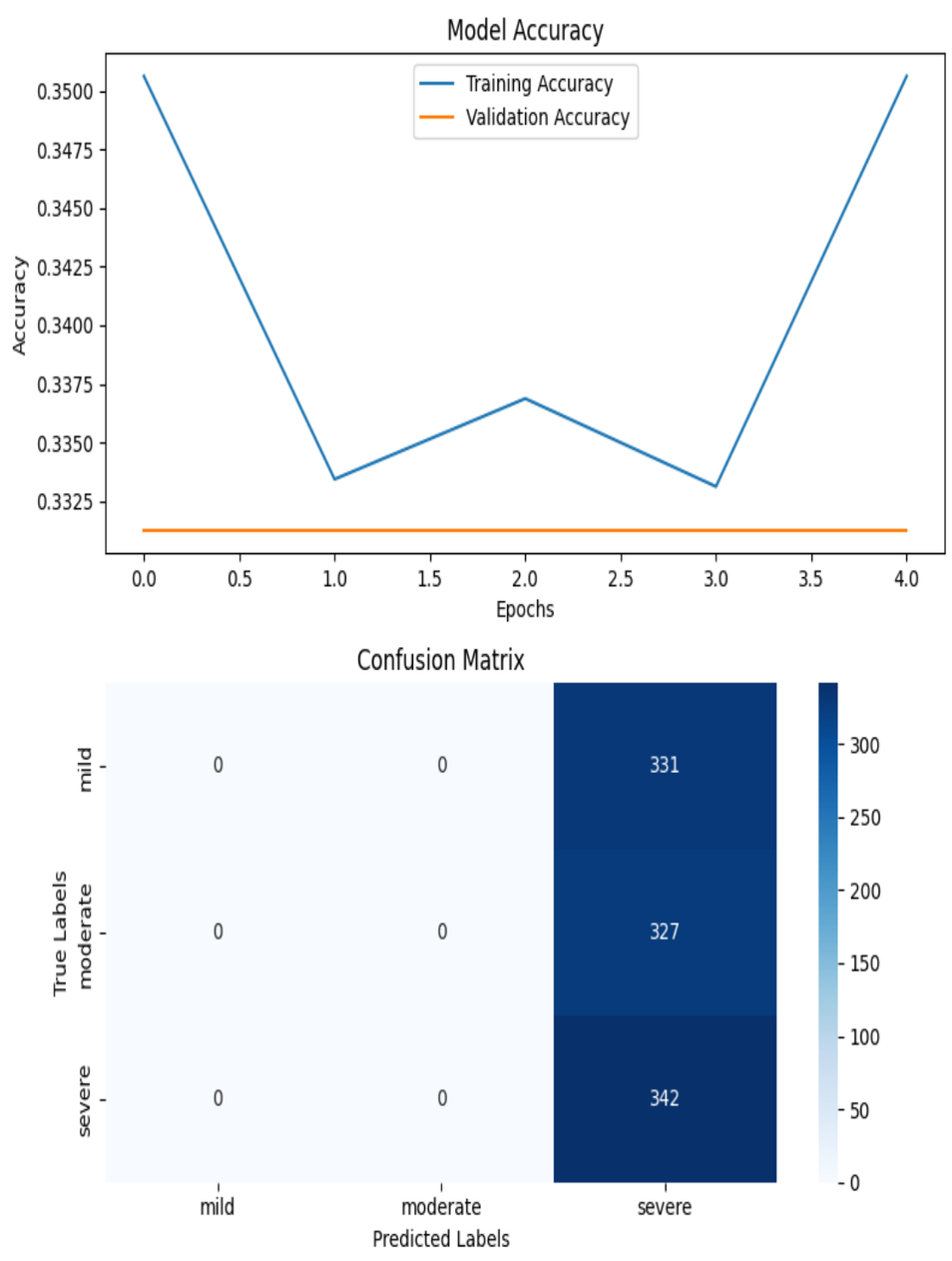
D. Results

The models were trained on the training set, with validation performed on the validation set during training. The best weights from each model were saved and used to predict labels on the test data. Training and testing accuracy and loss curves for learning rates of 2e-5, 5e-5, and 8e-5 are shown in Figures 3 and 4. All models demonstrated good performance, with quick convergence and stable training curves. Specifically, at a learning rate of 8e-5, ESG, ESD, and XDL models converged to the highest validation accuracy within just two epochs, whereas ABV required more epochs for smooth convergence.In terms of F1 score, ESG and ABV achieved the best score of 89%, showing strong contextual understanding of short tweets. However, ESG demonstrated a faster training time, taking an average of 130 seconds per epoch, compared to ABV's 410 seconds. XDL, the fastest model in terms of training time, took only 75 seconds per epoch and achieved an F1 score of 88%, making it ideal for resource-constrained environments.Confusion matrices (Figure 3) show that ESG achieved the highest number of correct predictions (9753), followed by ABV, ESD, and XDL. The lowest correct predictions were made by XDL with a 2e-5 learning rate. Micro-average scores in Table 6 demonstrate that ABV and ESG outperformed Distil Bert in terms of F1 score, even though Distil Bert has more parameters (68 million).

E. Optimization of Model

Model optimization becomes crucial when deploying models on devices with limited computational power, such as mobile phones or embedded systems. Several optimization techniques, like pruning and quantization, are explored to make the models suitable for such environments.

- **Pruning** reduces the number of parameters by setting weights to zero during training, resulting in a sparser model.
- **Quantization** reduces the precision of model weights, making the model smaller and faster without significant loss in accuracy.




```
Classification Report:
              precision    recall  f1-score   support
mild           0.000000  0.000000  0.000000   331.000
moderate       0.000000  0.000000  0.000000   327.000
severe         0.342000  1.000000  0.509687   342.000
accuracy       0.342000  0.342000  0.342000     0.342
macro avg      0.114000  0.333333  0.169896  1000.000
weighted avg   0.116964  0.342000  0.174313  1000.000
Confusion Matrix:
          mild  moderate  severe
mild         0         0     331
moderate     0         0     327
severe       0         0     342
```

F. Post-Training Quantization of Model

To optimize ESG for deployment on low-power devices, post-training quantization is applied using TensorFlow Lite. Initially, ESG is trained with 32-bit floating point precision, which results in a model size of 57 MB. After quantization, the model is reduced to 27 MB with minimal impact on classification performance. The F1 score slightly decreases from 89% to 88%, but this trade-off is acceptable considering the significant reduction in model size.Quantization reduces memory usage by 50%, making the model more suitable for deployment on devices with limited resources, such as smartphones or microcontrollers. The workflow of ESG quantization and evaluation is shown in Figure 3.In conclusion, ESG offers the best balance between speed, performance, and resource efficiency, making it the optimal choice for real-time depression classification tasks on devices with varying computational capabilities.

CONCLUSION

This study investigates depression intensity classification using Twitter data by evaluating four small transformer-based language models. A custom corpus of 73,355 tweets, categorized into three intensity levels—'severe,' 'moderate,' and 'mild'—is created for this task. Through a comprehensive evaluation using transfer learning and downstream fine-tuning, we demonstrate that ESG emerges as the most effective model, achieving a high F1 score of 89% in a short training time of 130 seconds per epoch. Notably, ESG converges efficiently within just two epochs when trained with a higher learning rate of 8e-5. ABV, while slower, excels in terms of accuracy.The study also compares the performance of these small models with Distil Bert, a larger model with 67 million parameters, revealing that the smaller models deliver comparable performance despite having fewer parameters. This finding challenges the conventional notion that larger models are always superior, offering a compelling case for using smaller models in real-time applications where computational resources are limited.Additionally, the quantization of the best-performing model, ESG, is explored, successfully reducing its size by 50% without significantly affecting its performance. This optimization makes ESG suitable for deployment on devices with constrained hardware, such as smartphones or embedded systems. The ability to accurately classify depression intensity plays a crucial role in the early detection of mental health issues, potentially preventing severe

outcomes like suicide.Looking ahead, future research may involve comparing the performance of the proposed model with a weighted ensemble of traditional machine learning algorithms, such as Naive Bayes, Logistic Regression, and Support Vector Machines, to explore further improvements in training time without sacrificing performance. The use of the Receiver Operating Characteristics (ROC) curve could offer additional insights into model performance, complementing metrics such as accuracy, recall, F1 score, and precision. Furthermore, while the current model is trained on short tweets, there is potential to extend this research by using longer text snippets, such as those from Reddit, to create a more generalized model capable of detecting depression across varied textual formats.

REFERENCES

1]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[2].Zhang, Y., Zhao, L., & Li, Q. (2020). Transformer-based approaches for sentiment analysis: A comprehensive review. *Neurocomputing, 417*, 134–146. https://doi.org/10.1016/j.neucom.2020.07.024

[3].Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *Proceedings of ACL*, 328–339. https://doi.org/10.18653/v1/P18-1031

[4].Radford, A., Narasimhan, K., & Salimans, T. (2018). Improving language understanding by generative pre-training. *OpenAI Blog*. https://blog.openai.com/language-unsupervised/

[5].Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Proceedings of NeurIPS*, 30, 5998–6008. https://arxiv.org/abs/1706.03762

[6].Guo, H., Zhang, L[., & Jiang, L. (2021). Depression detection on social media: A survey. *IEEE Access, 9*, 128177–128190. https://doi.org/10.1109/ACCESS.2021.3108243

[7].Albrecht, R., & Martin, J. (2019). Detecting depression in tweets using transformers. *Proceedings of the IEEE International Conference on Big Data*, 3376–3383. https://doi.org/10.1109/BigData47090.2019.9006369

[8].Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *Proceedings of ACL*, 328–339. https://doi.org/10.18653/v1/P18-1031

[9].Radford, A., Narasimhan, K., & Salimans, T. (2018). Improving language understanding by generative pre-training. *OpenAI Blog*. https://blog.openai.com/language-unsupervised/

[10.].Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Proceedings of NeurIPS*, 30, 5998–6008. https://arxiv.org/abs/1706.03762