IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 14 October 2022, accepted 14 November 2022, date of publication 17 November 2022,
date of current version 15 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3223049

RESEARCH ARTICLE

Depression Classification From Tweets Using Small Deep Transfer Learning Language Models

MUHAMMAD RIZWAN[1,5], MUHAMMAD FAHEEM MUSHTAQ[1], UROOJ AKRAM[1], ARIF MEHMOOD[2], IMRAN ASHRAF[3], AND BENJAMÍN SAHELICES[4]

[1]Department of Artificial Intelligence, Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan
[2]Department of Computer Science and Information Technology, Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan
[3]Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, South Korea
[4]Department of Informatics, University of Valladolid, 47002 Valladolid, Spain
[5]Department of Information Technology, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan 64200, Pakistan

Corresponding authors: Imran Ashraf (ashrafimran@live.com) and Benjamín Sahelices (benjamin.sahelices@uva.es)

This work was supported in part by the Department of Informatics, University of Valladolid, Spain; in part by the Spanish Ministry of Economy and Competitiveness through Feder Funds under Grant TEC2017-84321-C4-2-R; in part by MINECO/AEI/ERDF (EU) under Grant PID2019-105660RB-C21 / AEI / 10.13039/501100011033; in part by the Aragón Government under Grant T58_20R research group; and in part by the Construyendo Europa desde Aragón under Grant ERDF 2014-2020.

ABSTRACT Depression detection from social media texts such as Tweets or Facebook comments could be very beneficial as early detection of depression may even avoid extreme consequences of long-term depression i.e. suicide. In this study, depression intensity classification is performed using a labeled Twitter dataset. Further, this study makes a detailed performance evaluation of four transformer-based pre-trained small language models, particularly those having less than 15 million tunable parameters i.e. Electra Small Generator (ESG), Electra Small Discriminator (ESD), XtremeDistil-L6 (XDL) and Albert Base V2 (ABV) for classification of depression intensity using Tweets. The models are fine-tuned to get the best performance by applying different hyperparameters. The models are tested by classification of depression intensity of labeled tweets for three label classes i.e. 'severe', 'moderate', and 'mild' by downstream fine-tuning the parameters. Evaluation metrics such as accuracy, F1, precision, recall, and specificity are calculated to evaluate the performance of the models. Comparative analysis of these models is also done with a moderately larger model i.e. DistilBert which has 67 million tunable parameters for the same task with the same experimental settings. Results indicate that ESG outperforms all other models including DistilBert due to its better deep contextualized text representation as it gets the best F1 score of 89% with comparatively less training time. Further optimization of ESG is also proposed to make it suitable for low-powered devices. This study helps to achieve better classification performance of depression detection as well as to choose the best language model in terms of performance and less training time for Twitter-related downstream NLP tasks.

INDEX TERMS Depression classification, transfer learning, transformer language models, public health.

I. INTRODUCTION

Depression has become the leading global mental disorder, particularly in the context of the COVID-19 outbreak. It is a mood disorder, which often triggers feelings of sadness or anger and affects the daily activities of individuals [1]. According to the world health organization, there are over 260 million patients with depression and it is the leading cause of non-fatal health issues globally [2]. Alarmingly, it is thought to be the second leading cause of suicide among young people. Every year more than one million people die because of depressional suicide. A lack of interest in social activities and everyday tasks of life can lead to many physical and mental problems like weight loss/gain, disturbed sleeping patterns, lack of concentration, and feelings of depravity and guilt, among others, [3]. The importance of an automatic approach to detect depression and determining its severity is very important. The early detection of depression-related symptoms could help to prevent suicides and deaths from depression [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Mervat Adib Bamiah.

Social media posts have been used for several disease detection tasks recently. The early warning symptoms of cancer can be detected from online activities of the people [5]. Depression and post-traumatic stress disorder from social media texts have been investigated in recent studies. The methods for depression classification can have a huge impact on the health of the general public. People use social media to express feelings of depression and loneliness [6], [7]. Studies show that the young generation is more likely to use social media to express suicidal thoughts than their parents or friends [8].

Keeping in view these repercussions, the primary objective of this study is to find the best language model among the transformer encoder language models with a smaller number of trainable parameters. The goal is to obtain better performance and less training time which can predict the intensity of depression in short text similar to tweets. This kind of study has not been done on the dataset of labeled tweets used in this work, as the dataset is newly created. Moreover, the majority of the work on depression detection focused on binary classification where the people suffering from depression are classified. However, this study takes this process one step further and focuses on the multi-class classification, and splits the victims of depression into different classes regarding the intensity of depression.

From this perspective, this study makes several unique contributions. This study evaluates the performance of four language models by classification of labeled tweets into depression severity class labels which are 'severe', 'moderate', and 'mild'. Four small (having less than 15 million tunable parameters) transformer-based language models are used for depression intensity classification through transfer learning of labeled tweets which include Electra small generator (ESG), Electra small discriminator (ESD), XtremeDistil-L6 (XDL), and Albert base v2 (ABV). Further, performance evaluation of the four language models has been done in terms of the training time as well as F1 and specificity. We also compare the performance of the models with DistilBert which is a larger model as compared to ESG, ESD, XDL, and ABV. It has 67 million tunable parameters and is a much larger model compared to other models. Comparison with a larger model further validates the study to get reliable results. Moreover, the models are analyzed in terms of early over-fitting concerning the F1 score. All four language models have less than 15 million parameters which makes them suitable for transfer learning and tuning with relatively less computational complexity. All models are trained downstream to fine-tune them using different hyperparameters and are further evaluated with different evaluation metrics.

The rest of the paper is organized into five sections. Section II discusses several important research works related to this study. The proposed approach is presented and explained in Section III. The experimental setup is elaborated in Section IV which is followed by the discussion of the results. In the end, the study is concluded.

II. RELATED WORK

In this section, we review the latest research regarding the use of transformer-based language models for natural language processing (NLP) applications such as text classification, named entity recognition (NER), as well as different disease predictions using social media texts which are helpful to find research gaps in the existing literature and contemporary state-of-the-art approaches.

Several investigations have been conducted regarding different NLP tasks with transformer-based language models. Bidirectional encoder representation from Transformers (BERT) is a very famous language model that has been used in several state-of-the-art studies for obtaining contextualized embedding of textual data for different NLP tasks. For example, researchers have been using BERT for deep contextualized embedding to perform sentiment analysis by using the downstream fine-tuning of the parameters and attention heads. In addition, deep transfer learning is also used which takes the pre-trained transformer model and further fine-tunes it for specific tasks [9], [10], [11], [12]. BERT is also used in combination with other deep sequence models such as gated recurrent unit (GRU) and long short-term memory (LSTM), etc. Other approaches also compared BERT-based sentiment analysis with lexicon-based study and BERT showed much better results than other models [13]. BERT is also studied in combination with convolution neural networks (CNN) which improved the performance as compared to the original BERT model [14]. Several studies used BERT for biomedical NER which has been done by combining the conditional random field (CRF) layer with multi-lingual BERT [15]. Similarly, Arabic biomedical NER has been done using a variant of BERT named AraBERT and multi-lingual BERT to obtain an F1 score of 85% [16]. In the same way, Chinese clinical NER is done using BERT with bidirectional LSTM and CRF in [17] and [18].

As BERT is a generic language model which can be used in different important NLP tasks such as question answering, text classification, and NER and sequence tagging, etc., different distilled versions of BERT have been introduced to reduce language models size [19], [20]. Distillation [21] is the process of training a small model as a 'student' which should mimic the maximum possible performance of a larger model as a 'teacher' like BERT. As the original BERT-based model has 110 million tunable parameters which are a bit large to be fine-tuned on a machine having limited computational resources, different distilled models based on BERT have also been proposed. For example, DistilBert [22] is a similar model which is smaller in terms of tunable parameters and faster in terms of training performance. Researchers used different techniques to distill the larger models to compress or reduce the parameters. The goal is to reduce the model by minimal compromise on the performance of the model [23], [24], [25], [26]. BERT distilled variants for multi-lingual text have also been proposed [27]. In [28], the authors proposed a model based on BERT and DenseNet which identifies multi-model tweets containing both image and text data

during the disaster. Regarding the combination of fine-tuning BERT with aspect-based sentiment analysis, a model for event detection is proposed using Twitter data in [29].

Several recent studies used social media data such as Twitter, Instagram, Reddit, etc. by applying pre-trained language models like BERT. For example, [30] used social media posts from Twitter and Facebook with BERT and deep learning models for analyzing attitudes towards COVID-19 vaccines. In another study, the authors classified garlic-related misinformation in the context of COVID-19 using different BERT-based pre-trained model variants using a large Twitter corpus [31]. The study [32] proposed a model named DICE which uses deep contextualized embedding of BERT in addition to a bidirectional LSTM network for sentiment detection. With the collective text and image features in Tweet using BERT fine-tuning and DesneNet, a model has been proposed to detect multi-model tweets in disaster [28]. A hybrid solution for event detection on Twitter is proposed in [29] by fine-tuning the BERT with aspect-based sentiment analysis.

Early depression detection using the beck depression inventory is carried out using the Reddit posts by applying BiLSTM and Albert language model in [32]. Using BERT and its three variants, a study was proposed to detect and classify social media toxicity using the Kaggle dataset [33]. Different variants of transformer language models are proposed for biomedical such as BioElectra, BioAlbert, and BioBert, etc., [34]. ELECTRA, BERT, and LSTM are combined for detecting suicide tendencies from social media text [35]. By combining the ELECTRA with LSTM a model was proposed for the emotional classification of Chinese text by fine-tuning the parameters with a softmax classification layer on top of the network [36]. BERT transfer learning in combination with CNN has been applied for the classification of Twitter posts containing both images and text in [37]. Similarly, a study used BERT feature embedding for binary classification of Twitter-based user depression [38]. For detecting depression in Arabic society, [39] proposed CarioDep which uses the BERT Arabic variants such as AraBERT and MARBert, etc.

Numerous studies have been conducted specifically for depression detection-based tasks using social media data by applying different transformer-based architectures. For example, using Twitter application programming interfaces (APIs), the depression-related tweets are collected and filtered by dividing Twitter users into 'diagnosed' vs 'control' in [40]. With the help of the geolocation field in tweet data, authors separated tweets concerning the country of origin, keeping the diagnosed vs control tweets separated for each country. They applied different existing machine learning and deep learning models to compare the results. Results show that BiLSTM-SELFA gives better performance and obtains up to 68% F1 score for the binary classification task. The study also analyzed the relationship between events like Christmas and COVID-19 with depression. A tweet monitoring framework is proposed to detect the user who is suspected to be at risk of depression using the social media data in [41].

The authors proposed a machine learning model to identify sadness among school students in [42]. The study showed that depression is the second stage of sadness, and often results from excessive stress and anxiety due to school workload. To find different health disorders, specifically depression, a qualitative analysis was performed in [43]. For data annotation, coding schemes of 6 resources were developed based on depression symptoms and psycho-social stress provided by different research articles. Studies [44], [45] use Latent Dirichlet allocation (LDA) to find depression among students. A large dataset of tweets is used for experiments using the newly proposed approach called auto-aggressive integrated moving average (ARIMA). Different depression and suicide-related trends and their corresponding deviations are also identified. To identify suicidal thoughts using Twitter data, the suicide artificial intelligence prediction heuristic (SAIPH) is proposed in [46]. The authors constructed different binary classification models for different use cases such as stress, insomnia, anxiety, loneliness, etc.

Analysis of the existing literature on depression detection and classification using social media data shows that predominantly such works focus on binary classification. The problem of multi-classification is rarely studied in the context of depression classification. Often the focus of the studies is to divide the victims into depressed and healthy subjects using tweets or short texts from Reddit and Facebook, etc. Therefore, this study focuses on the depression intensity classification where the intensity is categorized as 'severe', 'moderate', and 'mild'.

III. MATERIALS AND METHODS

This study presents an approach for depression intensity classification using tweets by employing four small transformer encoder-based language models. Performance evaluation of all models is carried out to quantify the best model with a higher F1 score and less training time. Figure 1 shows the workflow diagram of the study.

A. DATASET

For experiments, depression-related tweets are extracted. For this purpose, tweets are gathered using Twitter public APIs by putting different depression-related hashtags as seed words. Previous studies show that users suffering from depression tend to show negative sentiments in tweets [47]. Further tweets are annotated using the Python libraries i.e. valence aware dictionary and sentiment reasoner (VADER) and TextBlob for calculating quantitative sentiment polarity and subjectivity scores of tweets. Tweets having less subjectivity are discarded and filtered out from the dataset to get opinionated tweets only. Table 1 shows a few sample tweets from the collected data along with their assigned labels.

Labels are assigned into three depression intensity classes i.e. 'mild', 'moderate', and 'severe' according to ICD-10 depression diagnostic criteria [48], [49].

$$D = \{t_i : \forall i \in \mathbb{Z} | t_1, t_2, \ldots, t_n\} \quad (1)$$

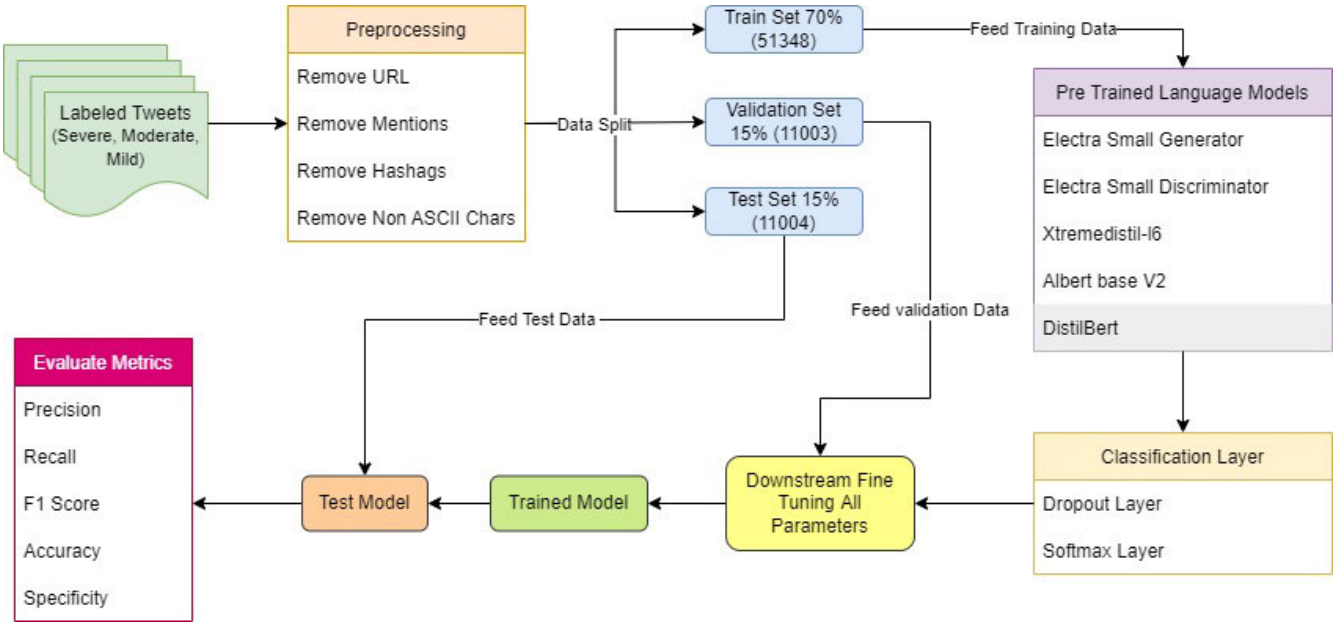

FIGURE 1. Overall architecture of depression intensity classification using deep transfer learning approach.

TABLE 1. Sample tweets with class labels.

S.No.	Tweet	Class label
1	I'm not saying isnt a problem... But based on streams takes the most "sick days" when not 'sick'... I GET IT THOUGH, SINCE ELEMENTARY SCHOOL..But its not something the working world calls 'sick' (though it should)Bad people cards!	Severe
2	Keep it open, keep it real! Let's chat guys. If u, or a man in your life, needs to chat... I'm here to sit in the shit with ya.	Moderate
3	I become my own hero as becoming reliant on others is only setting urself to fall again. If you won't help urself then don't expect anyone to do it for you	Mild
4	i wish i could believe in . i wish that i believed in something after this life. but i only believe in this one life. and its hard. and im not happy. and i cant currently make any changes. where does that leave me?	Severe
5	turn to friends and family members who make you feel loved and cared for. stop	Mild
6	you don't understand until you can't stand your own presence in an empty room.	Moderate

$$(S_b(t_i) > 0.5) \quad \forall D = D_a \tag{2}$$

$$A(D_a, SN()) = \begin{cases} Mild & if\ SN(t_i) < -0.1 \\ Moderate & if\ -0.1 \leq SN(t_i) \leq +0.3 \\ Severe & if\ SN(t_i) > +0.3 \end{cases} \tag{3}$$

where D shows the dataset containing the collected tweets. Each tweet t_i is used to compute quantitative subjectivity where $\{S_b(t_i) \in \mathbb{R} | 0 \leq S_b(t_i) \leq 1\}$ and $SN(t_i)$ is a function to compute quantitative sentiment score. Then labels are assigned based on $\{SN(t_i) \in \mathbb{R} | -1 \leq SN(t_i) \leq +1\}$.

Although several social media platforms have been used for depression analysis recently including Twitter, Reddit, Wiebo, etc., this study selects Twitter for two reasons. First, a predominantly large number of studies have used Twitter for data extraction and analysis for depression analysis. Twitter and Reddit have been the most famous platforms for depression-related machine learning and NLP problems [50] as compared to other social media platforms. Second, the number of users on the Twitter platform is high compared to other social media platforms that use the English language.

Table 2 shows the number of records for each class after the collected tweets are labeled using VADER. It indicates that the 'mild' and 'moderate' classes have almost a similar number of samples while the 'severe' class has a comparatively lower number of samples.

TABLE 2. Class label and corresponding tweet counts.

Label	Tweets count
Mild	29,931
Moderate	28,106
Severe	15,331

B. TRANSFORMER ENCODER-BASED PRE-TRAINED LANGUAGE MODELS

Figure 1 shows the architecture of the classification of depression through deep transfer learning using downstream fine-tuning of pre-trained language models. Predominantly, contemporary NLP applications and systems are using pre-trained language models with encoder transformers. The main task of these models is to train a large corpus that serves pre-trained startups to further fine-tune specific NLP tasks such as text classification and NER, etc. Fine-tuning of

these pre-trained models can be done by adding extra layers according to the nature of the NLP task it is trained for. Transformer-based models with the self-attention technique changed the deep contextual text representation for language models.

As the text is a sequence of characters or words, the natural fit for text data is sequence-based deep neural network models such as RNN, GRU, LSTM, etc. But training in parallel could benefit more in the case of the transformer as compared to RNN-based architecture as RNN feeds word by word into the network but the transformer-based model feeds the input text as a whole. Transformer models use a self-attention mechanism that covers the context of large sequences e.g. long sentences. Transformer neural network architecture has a major advantage in the parallelization of sequential data. The basic encoder-decoder structure in a transformer is the same as RNN or LSTM but the main difference in the transformer is that data can be processed in parallel which makes it possible to train on a large corpus. The positional encoder provides contextual information training and word vectors which are not available in static word embeddings like word2vec or global vectors (GloVe), etc. In each encoder block, we have a multi-head attention layer and a fully connected layer. To get the context of the word in a text document, the embedded input is used to get the vector shape of the word. Two of the main components of the decoder block are the same as the ones used in the encoder blocks. Each word in the text document has a self-attention block which shows how many words are related to each other. The output in the form of attention is sent to the next feed-forward layer, linear layer, and softmax probabilities.

1) BERT

The BERT model has proved to be very helpful in complex tasks on natural language datasets such as sentiment classification and prediction of masked words. Its architecture is based on a stack of trained transformer encoders. The model can be generated by adding the context of specific words from the sentence or document. The BERT model helps to retain long-term dependency in sentences up to a maximum of 512 words by using the self-attention mechanism. 512 words of contextual ability are sufficient for most of the NLP tasks but sentences with more than 512 words may be truncated to train the model. BERT uses a loss function based on the score of masked word prediction to get the bidirectional context of masked words. Further, BERT also uses next-sentence prediction during training. This makes it capable of identifying two words as identical or not in terms of their respective context. Natural language inference and semantic text similarity are expected to be improved with the help of next-sentence prediction.

During the training phase, BERT gains an understanding of a token's context from both the left and right sides to get a deep contextualized representation of the text document. Translation, question answering, text classification, and text summarization are just some of the practical use cases of BERT and BERT-based models. A contextual language understanding is required for all of the mentioned examples. BERT can be trained in English or multiple languages. Downstream training of the model can make it better for a specific dataset. The training of BERT can be done in two parts; in the first part language context is identified using a self-attention mechanism and in the second part, fine-tuning of tunable parameters can be done to get a high score prediction. Pre-training in BERT makes it fit to learn the deep context of the word within sentences or paragraphs. Next, we discuss other BERT-based language models used in this study which follow the BERT architecture with further reduction in model size by using different distillation techniques e.g. DistilBert, Electra, Albert, and Xtremedistil, etc.

2) ELECTRA

Electra stands for 'efficiently learning an encoder that classifies token replacement accurately'. There are small and large versions of Electra available, but in this study, we only use Electra small which contains 13.5 million tunable parameters. Electra has been trained jointly with two models i.e. generator and discriminator. The generator is trained using a masked language model (MLM) by replacing random words with a mask to fine-tune the model for predicting the masked words [36], [51]. On the other hand, the discriminator is trained to identify which tokens match the original input from the generator samples.

3) XtremeDistil

There are two types of knowledge distillation, i.e. task specific distillation which models compression technique based on training, and task agnostic distillation. The former has the benefit that it only needs to be distilled once and can be used for any other NLP downstream fine-tuning tasks, but later has the advantage of comparatively high compression of the model. XDL uses the large BERT and Electra model as a teacher and a short version of MiniLM to mimic the student for its distillation tasks and uses the task agnostic distillation method. XDL has multiple variants available on 'huggingface' regarding encoder layers, hidden size, and attention heads. In this study, Xtremedistil-l6-h256-uncased has been used keeping the small model parameter in mind, as it only has 12.7 million tunable parameters [52].

4) ALBERT

Various variants of Albert are available on 'huggingface', i.e. Albert base, large, xlarge, and xxlarge, etc. A larger model tends to increase the number of tunable parameters. Although the largest Albert model xxlarge has lesser parameters as compared to BERT but computationally more expensive than BERT due to its bigger structure. Albert uses two model compression techniques to reduce parameters. First, it reduces the embedding matrix into relatively small matrices which split the hidden layer size from the embeddings matrix thus making it easier to increase the hidden layer size. Second, Albert shares the parameters of other layers [53]. For increasing the

TABLE 3. Models used in the study.

Model	Tunable parameter (millions)	Pretrained model download size	Huggingface link
albert-base-v2	11.6 M	63 MB	huggingface.co/albert-base-v2
electra-small-discriminator	13.5 M	55 MB	huggingface.co/google/electra-small-discriminator
electra-small-generator	13.5 M	70.4 MB	huggingface.co/google/electra-small-generator
xtremedistil-l6-h256-uncased	12.7 M	51 MB	huggingface.co/microsoft/xtremedistil-l6-h256-uncased

FIGURE 2. Architecture to classify depression intensity with pre-trained models.

performance of Albert, sentence order prediction (SOP) loss is used which gives better performance as compared to the next sentence prediction loss used in the BERT model.

C. PREPROCESSING OF TWEETS

To minimize the noise in data, pre-processing steps are essential for NLP-based tasks in general. Hashtags and universal resource locator (URL) are removed from the collected data. User identities that start with @ sign are also removed. Non-Ascii words have been replaced with white space.

IV. RESULTS AND DISCUSSIONS

A. TRAIN VALIDATION TEST SPLIT OF DATA

In this step, the labeled tweets are split into training, test, and validation sets to ensure an equal ratio of each class in all sets. For splitting, the Python Sklearn library is used and its 'train_test_split' function is used which supports the splitting of data in a stratified fashion.

B. EXPERIMENTAL SETUP

Table 3 shows four small models which have been selected for this study i.e. ESG, ESD, XDL, and ABV. These models are employed in this study and the performance of these models is further compared with the larger model DistilBert for the classification of tweets concerning three class labels 'severe', 'moderate', and 'mild'. Tweets are tokenized by a tokenizer provided for each respective model on 'HuggingFace' which maps tokens to their respective IDs. The maximum token length of a tweet within the dataset is found to be 62 so all tweets are padded to a fixed length of 64 tokens. The dataset is divided into three splits i.e. 70% for the training set having 51348 tweets, 15% for the test set consisting of 11004 tweets, and 15% for the validation set containing 11003 tweets.

For depression intensity detection, a classification layer is included at the end of each model as shown in Figure 2. The classification layer consists of a dropout layer with a softmax of size three which represent the intensity of depression. A dropout layer is added to avoid the early over-fitting of models during the training phase. Electra is pre-trained using Wikipedia and Bookcorpus [54]. A corresponding TensorFlow-based sequence classification interface e.g. 'TFElectraSequenceClassification' is used with each model. By feeding the training data to the pre-trained model, the classification layer with all tunable parameters is trained on specific depression intensity classification.

Fine-tuning is done using the hyperparameters as follows. The learning rate of three different values 2e-5, 5e-5, and 8e-5 is used to evaluate the models' performance on low, average, and high learning rates. Each experiment is done using the 'Adam' optimizer. The batch size is set to 64 for all experiments. The deep learning framework Tensorflow and Keras is used for model training. For optimal training, performance one cycle learning policy [55] is used which streamlines the best learning rate during the training process. In the first part of training, the learning rate gradually increases while it gradually decreases in the second part. Nvidia Tesla P100 GPU with 12 GB of RAM on an Ubuntu-based machine is used for all experiments. Table 4 shows the experimental settings for the models selected for depression intensity classification.

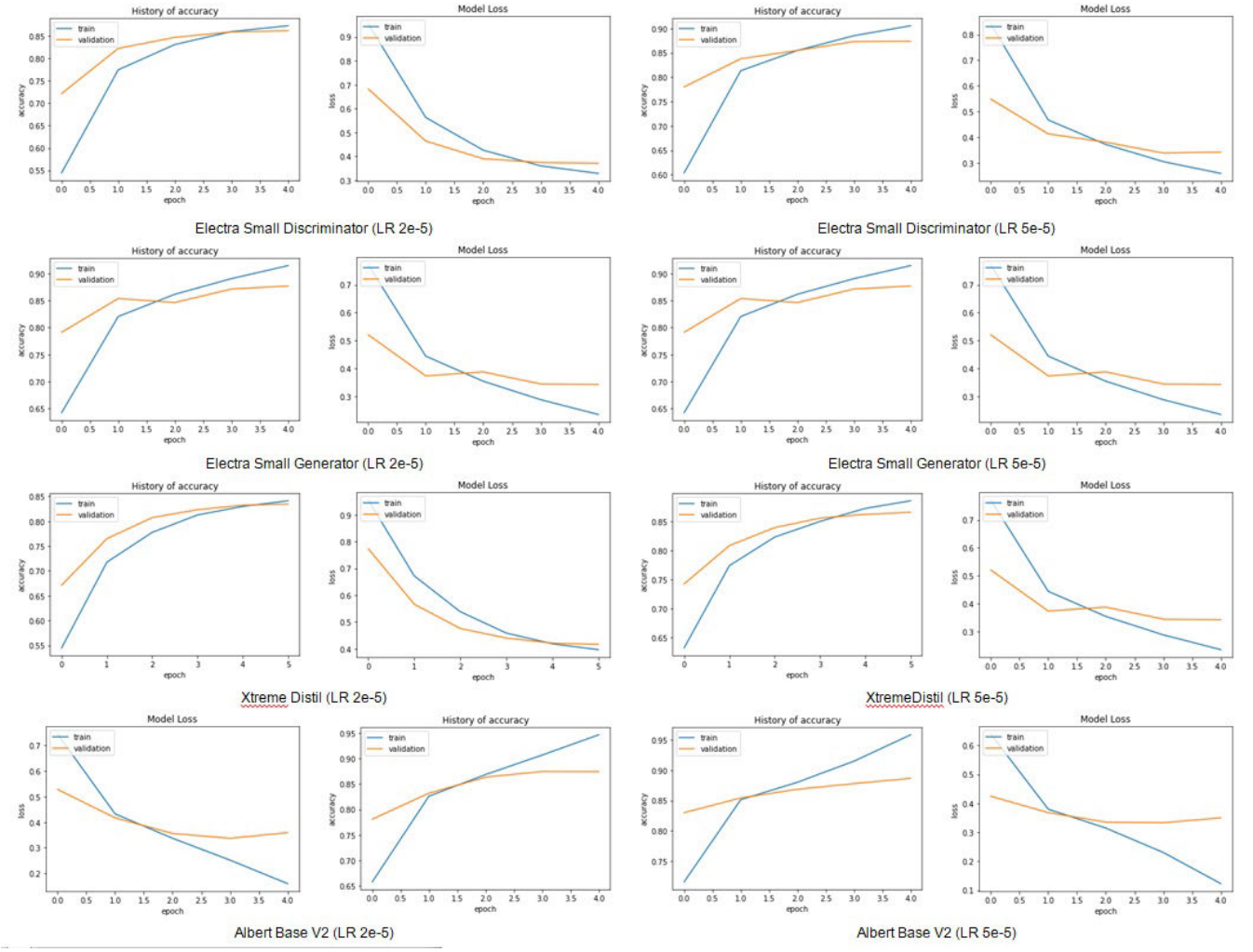

FIGURE 3. Loss and accuracy curves of each experiment at a learning rate of 5e-5 and 2e-5.

TABLE 4. Experimental settings.

Hyperparameter	Value (s)
Optimizer	Adam
Loss function	Categorical cross entropy
Learning Rate	2e-5, 5e-5, 8e-5
Batch Size	64
Tokens Length	64
Training With	One Cycle Policy

C. EVALUATION METRICS

The softmax layer predicts class labels by applying the trained model to the test dataset. A confusion matrix of 3×3 dimension is created for each experiment concerning the true label and predicted label. The confusion matrix provides accuracy for each class while also showing misclassification. The evaluation metrics such as accuracy, precision, recall, F1, and specificity are used to evaluate all models using scores from their corresponding confusion matrices. These parameters are used with the following equations

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{4}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{5}$$

$$Recall = \frac{TP}{(FN + TP)} \tag{6}$$

$$F1 = \frac{(2 \cdot Precision \cdot Recall)}{(Rrecision + Recall)} \tag{7}$$

$$Specificity = \frac{TN}{(TN + FP)} \tag{8}$$

where TP stands for true positive, FP for false positives, FN for false negative, and TN for true negative.

The 'severe' depression intensity class has a lower number of samples as compared to the two other labels, so F1 is an important metric due to the imbalanced nature of the dataset.

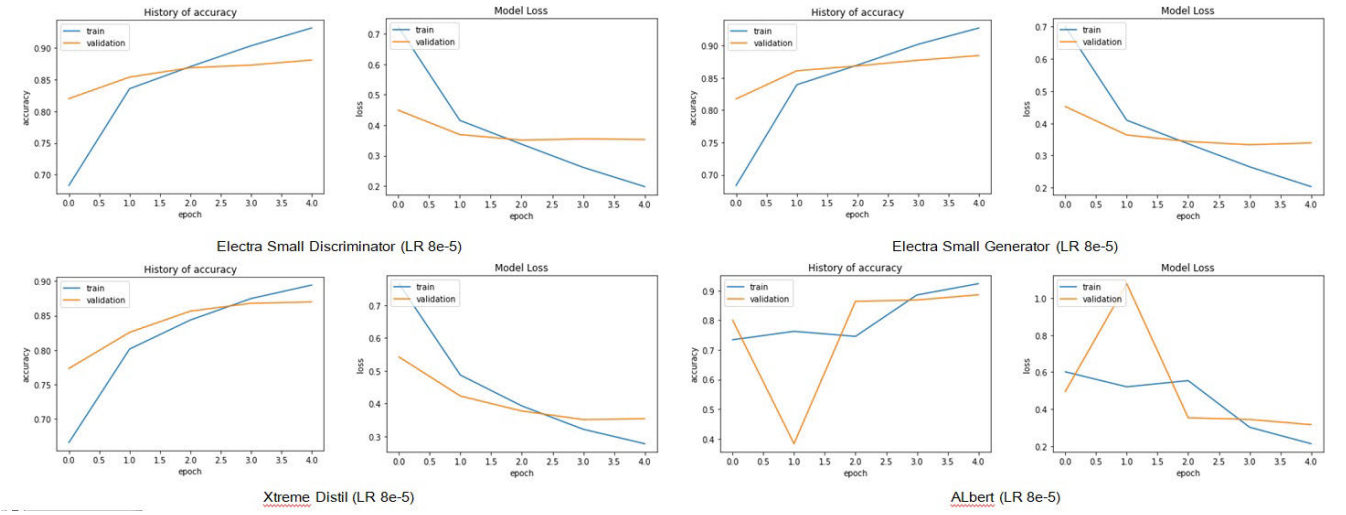

FIGURE 4. Loss and accuracy curves of each experiment at a learning rate of 8e-5.

Micro average scores are used to depict the performance of multi-class classification of depression detection.

D. RESULTS

Each encoder model is trained using a training dataset with validation performed on the validation dataset during the training. Moreover, the best model weights are saved and further used to predict the labels on test data to get samples of each model's performance for each experiment.

Training and testing accuracy and loss graphs are shown in Figures 3 and 4 for learning rates of 2e-5, 5e-5, and 8e-5, respectively. All encoder language model shows very good result in terms of training loss, validation & test accuracies, and speedy convergence in fewer training epochs. But specifically at the learning rate of 8e-5, ESG, ESD, and XDL converge quickly to the highest validation accuracy which is evident that at this learning rate only two epochs are enough to get the highest performance of downstream fine-tuning for classification. On the other hand, ABV is not much smooth at a learning rate of 8e-5 but rather shows a very smooth loss and validation curve at a lower learning rate of 5e-5 and 2e-5.

Regardless of test accuracy and training time, ESG and ABV obtain the best F1 score of 89% which exhibits the better capability of getting deep contextualized representation from short text-like tweets for the multi-class classification task. But if training is also in consideration then ESG is a clear winner as it possesses the 89% F1 in an average epoch training time of 130 seconds which is much lesser as compared to ABV which takes 410 seconds for epoch on the same machine and GPU environment with same hyperparameter settings of learning rate and batch size, etc. XDL, as its name indicates, is the fastest encoder model in the current study which only takes 75 seconds of training time for one epoch and achieves the F1 score of 88%. If a little compromise on classification performance is bearable, XDL is an exceptionally well model which gives appropriate accuracy and F1 with a very small training cost and competes with the advanced models in terms of capturing the sequence features using a contextualized representation of text in depression classification. It is also recommended for low parallel computing resources, as well as CPU-only machines. XDL and ESG yield relatively low F1 scores compared to ESG and ABV.

Figure 5 shows the confusion matrices of all the models used in this study to indicate their performance regarding the correct and wrong predictions at different learning rates. It shows that the best results are obtained at the learning rate of 8e-5 for all models regarding the number of correct predictions while the highest number of correct predictions are obtained by the ESG model, i.e. 9753 correct predictions followed by ABV, ESD, and XDL with 9750, 9733 and 9632 correct predictions, respectively. The lowest correct predictions of 9273 are made by the XDL model when trained using a 2e-5 learning rate.

Table 6 shows the results regarding the micro average. ABV and ESG outperform Distilbert regarding F1 score in the same experimental settings, even though DistilBert is a much larger model with 68 million parameters. Although ABV performs extremely well in terms of F1 and accuracy, ESG is advantageous and preferred over ABV because of its fast training time and early convergence. The highest F1 score is obtained by the ABV model which is 0.89 when a learning rate of 8e-5 is used and the same is true for its sensitivity.

Table 7 summarizes the model training time regarding the experiments. As previously discussed, if a little compromise can be made regarding accuracy, XDL is the best model as it requires a substantially shorter training time as compared to other models. As we deliberately selected small distilled language models, the parameter range of models varies within a narrow range of 12.7 million (XDL) to 13.5 million (ESG). Average training time ranges from 75 seconds(XDL) to

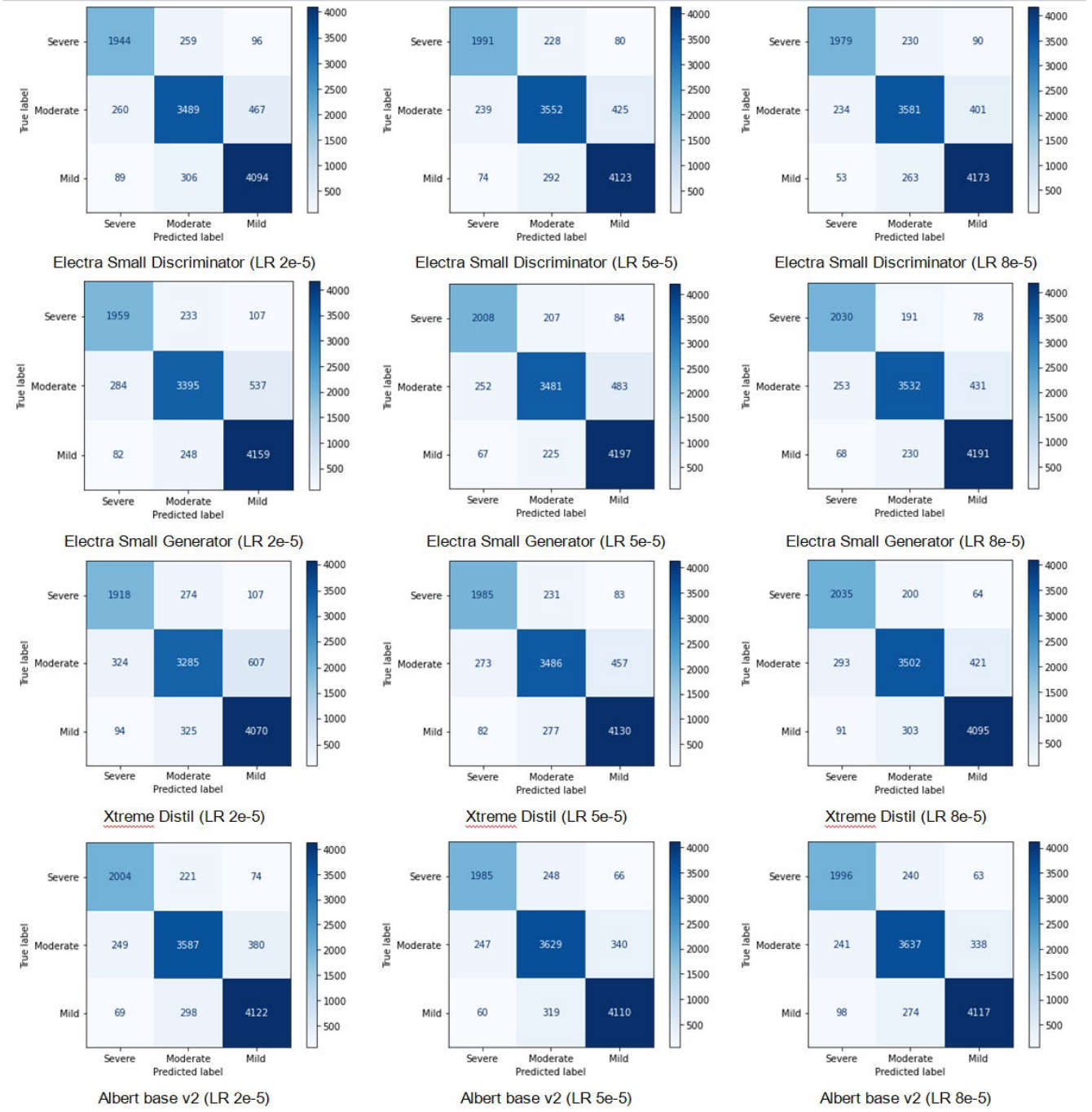

FIGURE 5. Confusion matrices for all models used for experiments.

410 seconds (ABV). Although models with a larger number of parameters tend to show high performance, this is not always the case. For example, in our case, ABV is the smallest model used for experiments regarding the number of parameters but still shows a high F1 score of 89%. Similarly, ABV is the slowest model to train due to its complexity, regardless of its small size.

E. OPTIMIZATION OF MODEL

Optimization of the model is required when we need to deploy a deep learning model in a device that has constraints in terms of computational power, memory usage, internet speed, etc., for example, mobile devices, IoT-based devices, and microcontroller devices. Another use-case of optimization is deploying a model in specially designed hardware.

TABLE 5. Class-wise evaluation score of all language models.

Model	LR	Class	Precision	Recall	Accuracy	F1	Specificity
Xtremedistil	8e-5	Severe	0.84	0.89	0.94	0.86	0.96
		Moderate	0.87	0.83	0.89	0.85	0.93
		Mild	0.89	0.91	0.92	0.90	0.93
Xtremedistil	5e-5	Severe	0.85	0.86	0.94	0.86	0.96
		Moderate	0.87	0.83	0.89	0.85	0.93
		Mild	0.88	0.92	0.92	0.90	0.92
Xtremedistil	2e-5	Severe	0.82	0.83	0.93	0.83	0.95
		Moderate	0.85	0.78	0.86	0.81	0.91
		Mild	0.85	0.91	0.90	0.84	0.89
Electra Small Discriminator	8e-5	Severe	0.87	0.86	0.94	0.87	0.97
		Moderate	0.88	0.85	0.90	0.86	0.93
		Mild	0.89	0.93	0.93	0.91	0.92
Electra Small Discriminator	5e-5	Severe	0.86	0.87	0.94	0.87	0.96
		Moderate	0.87	0.84	0.89	0.86	0.92
		Mild	0.89	0.92	0.92	0.90	0.92
Electra Small Discriminator	2e-5	Severe	0.85	0.85	0.94	0.85	0.96
		Moderate	0.86	0.83	0.88	0.84	0.92
		Mild	0.88	0.91	0.91	0.90	0.91
Electra Small Generator	8e-5	Severe	0.86	0.88	0.95	0.87	0.96
		Moderate	0.89	0.84	0.90	0.86	0.94
		Mild	0.89	0.93	0.93	0.91	0.92
Electra Small Generator	5e-5	Severe	0.86	0.87	0.94	0.87	0.96
		Moderate	0.89	0.83	0.89	0.86	0.94
		Mild	0.88	0.93	0.92	0.91	0.91
Electra Small Generator	2e-5	Severe	0.84	0.85	0.94	0.85	0.96
		Moderate	0.88	0.81	0.88	0.84	0.93
		Mild	0.87	0.93	0.91	0.90	0.90
Albert Base V2	8e-5	Severe	0.85	0.87	0.94	0.86	0.96
		Moderate	0.88	0.86	0.90	0.87	0.92
		Mild	0.91	0.92	0.93	0.91	0.94
Albert Base V2	5e-5	Severe	0.87	0.86	0.94	0.86	0.96
		Moderate	0.86	0.86	0.90	0.86	0.92
		Mild	0.91	0.92	0.93	0.91	0.94
Albert Base V2	2e-5	Severe	0.86	0.87	0.94	0.87	0.96
		Moderate	0.87	0.85	0.90	0.86	0.85
		Mild	0.90	0.92	0.93	0.91	0.93

TABLE 6. Micro average score of all language models.

Model	LR	Precision (Micro Avg.)	Recall / Sensitivity (Micro Avg.)	Accuracy (Micro Avg.)	F1 score (Micro Avg.)	Specificity (Micro Avg.)
Xtremedistil	8e-5	0.88	0.88	0.92	0.88	0.94
Xtremedistil	5e-5	0.87	0.87	0.92	0.87	0.94
Xtremedistil	2e-5	0.84	0.84	0.90	0.84	0.92
Electra Small Discriminator	8e-5	0.88	0.88	0.92	0.88	0.94
Electra Small Discriminator	5e-5	0.88	0.88	0.92	0.88	0.94
Electra Small Discriminator	2e-5	0.87	0.87	0.91	0.87	0.93
Electra Small Generator	8e-5	0.89	0.89	0.92	0.89	0.94
Electra Small Generator	5e-5	0.88	0.88	0.92	0.88	0.94
Electra Small Generator	2e-5	0.86	0.86	0.91	0.86	0.93
ALbert Base v2	8e-5	0.89	0.89	0.92	0.89	0.94
ALbert Base v2	5e-5	0.88	0.88	0.92	0.88	0.94
ALbert Base v2	2e-5	0.88	0.88	0.92	0.88	0.94
Distilbert uncased	2e-5	0.88	0.88	0.92	0.88	0.94

TABLE 7. Training time of the models.

Model	Avg. training time per epoch (seconds)
Albert-base-v2	410 seconds
Electra-small-discriminator	130 seconds
Electra-small-generator	130 seconds
xtremedistil-l6-h256-uncased	75 seconds

There are multiple types of optimization techniques proposed in contemporary research. For example, pruning techniques are used to minimize the number of parameters in the model. Model pruning is a compression technique in which model weights are converted to zero during the training phase to increase the sparsity in the model. A sparse model is easy to compress to further minimize latency in the network. The other technique is called quantization in which the model uses approximate lower precision of floating point weights during deployment. Quantization significantly reduces deployment model size which makes them suitable

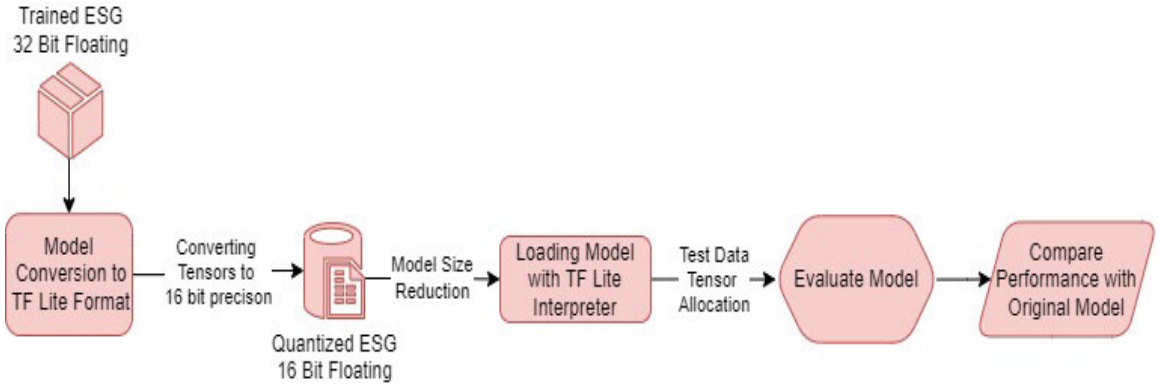

FIGURE 6. ESG trained model quantization and evaluation workflow.

for smartphones and micro-controllers [56], [57]. There are two types of quantization i.e. post-training quantization and training-aware quantization. The former is applied after training the model and the latter is applied during the training phase.

F. POST TRAINING QUANTIZATION OF MODEL

Quantization is the process of low-precision approximation floating point numbers regarding neural network weights in the form of tensors to significantly reduce the model size. So to further optimize the best-performing model in this study which is ESG, post-training quantization is proposed to reduce the model size so that it may be deployed in mobile devices as well or in a client-server architecture with low latency. ESG is already trained and fine-tuned for the classification of depression but is only suitable for deploying it on desktop-based systems. This study also aims to produce a lighter version of the model using the post-training quantization technique which can also be deployed on embedded devices as well as smartphones. Tensorflow (TF) Lite has been used for the quantization process [58]. TF Lite is specially designed for compressing deep learning models with different model optimization techniques so that models can be easily deployed on small memory-embedded devices.

ESG is trained with 32-bit floating point precision and weights of the trained model are stored in the same format and quite large for embedded and lower power devices in terms of deploying them in the main memory of low powered devices [59], [60]. Further during prediction, the model also needs to perform intensive floating point calculation which is also impractical for low-powered devices. The trained ESG for depression classification is 57 MB with the full precision of a 32-bit floating point. Our aim is to quantize the model without compromising the classification performance of ESG. The workflow of quantization of ESG and evaluation are shown in Figure 6.

ESG architecture mainly consists of two layers i.e. multi-head attention layer and a classification layer of simple perceptron followed by a softmax layer. At the attention layer, the attention matrix is calculated by the dot product of the queries matrix and key matrix. The most expensive calculation is the matrix product of multi-head attention to the classification layer and this is optimized by quantization of all weights from 32-bit floating point numbers to 16-bit floating point numbers. It helps to reduce the model size as well as the cost of classifying new instances. We chose 16-bit floating quantization because it is well-suited for GPU-based smartphones. The 8-bit integer quantization option is also viable but it does not support all kinds of hardware.

It is evident from the experiments that proposed optimization through quantization saves around 50% memory by reducing the model size to 27MB which is almost half of the original size of the model. The compression ratio of the model is 2. It also reduces the computational cost of prediction and maintains almost the same classification performance in terms of accuracy of 92% and F1 score of 88% (slight reduction) in comparison to the original 32-bit floating precision trained model. Accuracy remains the same but a slight reduction in the F1 score is observed which is insignificant keeping in mind the model achieved compression ratio.

V. CONCLUSION

This study performs depression intensity classification using Twitter data by performing experiments on four small transformer-based language models. A comprehensive evaluation of these models is performed using transfer learning and downstream fine-tuning for multi-class classification of depression intensity. A dedicated corpus of 73355 tweets is created for experiments, comprising three levels of intensity, i.e., 'severe', 'moderate', and 'mild'. ESG proves to be the most effective model and outperforms other models in terms of a high classification score regarding F1 of 89% in a relatively short training time which is 130 seconds per epoch. In addition, it can easily converge with two epochs with a little higher learning rate of 8e-5. ABV is the best-performing model in terms of the highest accuracy. Further, the performance of transformer models with less than 15 million parameters is compared with the advanced model DistilBert with 67 million parameters. The study shows very interesting results that the performance of small language

models is very much comparable to DistilBert which is a much larger model in terms of tunable parameters. This study provides the impactful foundation for the choice of small language models for the classification of tweets in general and depression classification in specific. This study also helps researchers and data scientists to choose the best small language models which give sufficiently good performance in less training time. Further quantization of the best performing model i.e. ESG is proposed which successfully reduces the model size to half of the original size with an insignificant reduction in accuracy and F1 score. Quantization of the model enables it to be deployed on constrained devices with low hardware resources. Moreover, accurate depression intensity classification helps early detection of depression to avoid the worst-case scenario of suicide.

In the future, we compare the performance of our proposed model with a weighted ensemble of soft voting of different conventional machine learning algorithms such as Naive Bayes, Logistic Regression, Support Vector Machine, etc. to further seek the best model with the same performance but a shorter training time. The evaluation metric area of a Receiver Operating Characteristics (ROC) curve shall also be used to more precisely observe the performance of models in addition to accuracy, recall, F1, and precision. As we trained the model using tweets that consist of short words but the trained model might not be suitable to predict depression intensity in a longer snippet of text. In the future, the model may be trained using Reddit data to make it more generalized for the prediction of depression in shorter and longer text.

REFERENCES

[1] C. Jiang, Y. Li, Y. Tang, and C. Guan, "Enhancing EEG-based classification of depression patients using spatial information," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 566–575, 2021.

[2] (2021). *Depression*. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression

[3] S. G. Burdisso, M. Errecalde, and M. Montes-y-Gómez, "A text classification framework for simple and effective early depression detection over social media streams," *Expert Syst. Appl.*, vol. 133, pp. 182–197, Nov. 2019.

[4] World Health Organzation. (2019). *Suicide Data*. Accessed: Jan. 2, 2022. [Online]. Available: https://www.who.int/teams/mental-health-and-substance-use/data-research/suicide-data

[5] J. Paparrizos, R. W. White, and E. Horvitz, "Screening for pancreatic adenocarcinoma using signals from web search logs: Feasibility study and results," *J. Oncol. Pract.*, vol. 12, no. 8, pp. 737–744, Aug. 2016.

[6] G. Coppersmith, C. Harman, and M. Dredze, "Measuring post traumatic stress disorder in Twitter," in *Proc. 8th Int. AAAI Conf. Weblogs Social Media*, 2014, pp. 579–582.

[7] M. Nadeem, "Identifying depression on Twitter," 2016, *arXiv:1607.07384*.

[8] A. John, A. C. Glendenning, A. Marchant, P. Montgomery, A. Stewart, S. Wood, K. Lloyd, and K. Hawton, "Self-harm, suicidal behaviours, and cyberbullying in children and young people: Systematic review," *J. Med. Internet Res.*, vol. 20, no. 4, p. e129, Apr. 2018.

[9] D. Deepa and A. Tamilarasi, "Bidirectional encoder representations from transformers (BERT) language model for sentiment analysis task," *Turkish J. Comput. Mathem. Educ.*, vol. 12, no. 7, pp. 1708–1721, 2021.

[10] J. Zheng, J. Wang, Y. Ren, and Z. Yang, "Chinese sentiment analysis of online education and internet buzzwords based on BERT," *J. Phys., Conf. Ser.*, vol. 1631, no. 1, Sep. 2020, Art. no. 012034.

[11] L. Zhao, L. Li, X. Zheng, and J. Zhang, "A BERT based sentiment analysis and key entity detection approach for online financial texts," in *Proc. IEEE 24th Int. Conf. Comput. Supported Cooperat. Work Design (CSCWD)*, May 2021, pp. 1233–1238.

[12] M. Pota, M. Ventura, R. Catelli, and M. Esposito, "An effective BERT-based pipeline for Twitter sentiment analysis: A case study in Italian," *Sensors*, vol. 21, no. 1, p. 133, 2021.

[13] R. Catelli, S. Pelosi, and M. Esposito, "Lexicon-based vs. bert-based sentiment analysis: A comparative study in Italian," *Electronics*, vol. 11, no. 3, p. 374, Jan. 2022.

[14] J. Dong, F. He, Y. Guo, and H. Zhang, "A commodity review sentiment analysis based on BERT-CNN model," in *Proc. 5th Int. Conf. Comput. Commun. Syst. (ICCCS)*, May 2020, pp. 143–147.

[15] K. Hakala and S. Pyysalo, "Biomedical named entity recognition with multilingual BERT," in *Proc. 5th Workshop BioNLP Open Shared Tasks*, 2019, pp. 56–61.

[16] N. Boudjellal, H. Zhang, A. Khan, A. Ahmad, R. Naseem, J. Shang, and L. Dai, "ABioNER: A BERT-based model for Arabic biomedical named-entity recognition," *Complexity*, vol. 2021, pp. 1–6, Mar. 2021.

[17] X. Li, H. Zhang, and X.-H. Zhou, "Chinese clinical named entity recognition with variant neural structures based on BERT methods," *J. Biomed. Informat.*, vol. 107, Jul. 2020, Art. no. 103422.

[18] Y. Chang, L. Kong, K. Jia, and Q. Meng, "Chinese named entity recognition method based on BERT," in *Proc. IEEE Int. Conf. Data Sci. Comput. Appl. (ICDSCA)*, Oct. 2021, pp. 294–299.

[19] N. Jalal, A. Mehmood, G. S. Choi, and I. Ashraf, "A novel improved random forest for text classification using feature ranking and optimal number of trees," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 6, pp. 2733–2742, Jun. 2022.

[20] V. Rupapara, F. Rustam, A. Amaar, P. B. Washington, E. Lee, and I. Ashraf, "Deepfake tweets classification using stacked bi-LSTM and words embedding," *PeerJ Comput. Sci.*, vol. 7, Oct. 2021, Art. no. e745.

[21] C. Wu, F. Wu, and Y. Huang, "One teacher is enough? Pre-trained language model distillation from multiple teachers," 2021, *arXiv:2106.01023*.

[22] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.

[23] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "TinyBERT: Distilling BERT for natural language understanding," 2019, *arXiv:1909.10351*.

[24] W. Zhang, L. Hou, Y. Yin, L. Shang, X. Chen, X. Jiang, and Q. Liu, "TernaryBERT: Distillation-aware ultra-low bit BERT," 2020, *arXiv:2009.12812*.

[25] Y. Xu, X. Qiu, L. Zhou, and X. Huang, "Improving BERT fine-tuning via self-ensemble and self-distillation," 2020, *arXiv:2002.10345*.

[26] S. Sun, Y. Cheng, Z. Gan, and J. Liu, "Patient knowledge distillation for BERT model compression," 2019, *arXiv:1908.09355*.

[27] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "LightMBERT: A simple yet effective method for multilingual BERT distillation," 2021, *arXiv:2103.06418*.

[28] S. Madichetty, S. Muthukumarasamy, and P. Jayadev, "Multi-modal classification of Twitter data during disasters for humanitarian response," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 11, pp. 10223–10237, Nov. 2021.

[29] M. M. Rahimi, E. Naghizade, M. Stevenson, and S. Winter, "Service quality monitoring in confined spaces through mining Twitter data," *J. Spatial Inf. Sci.*, no. 21, pp. 229–261, Dec. 2020.

[30] A. Hussain, A. Tahir, Z. Hussain, Z. Sheikh, M. Gogate, K. Dashtipour, A. Ali, and A. Sheikh, "Artificial intelligence–enabled analysis of public attitudes on Facebook and Twitter toward COVID-19 vaccines in the United Kingdom and the United States: Observational study," *J. Med. Internet Res.*, vol. 23, no. 4, Apr. 2021, Art. no. e26627.

[31] M. G. Kim, M. Kim, J. H. Kim, and K. Kim, "Fine-tuning BERT models to classify misinformation on garlic and COVID-19 on Twitter," *Int. J. Environ. Res. Public Health*, vol. 19, no. 9, p. 5126, Apr. 2022.

[32] U. Naseem, I. Razzak, K. Musial, and M. Imran, "Transformer based deep intelligent contextual embedding for Twitter sentiment analysis," *Future Gener. Comput. Syst.*, vol. 113, pp. 58–69, Dec. 2020.

[33] H. Fan, W. Du, A. Dahou, A. A. Ewees, D. Yousri, M. A. Elaziz, A. H. Elsheikh, L. Abualigah, and M. A. A. Al-Qaness, "Social media toxicity classification using deep learning: Real-world application U.K. Brexit," *Electronics*, vol. 10, no. 11, p. 1332, Jun. 2021.

[34] K. S. Kalyan, A. Rajasekharan, and S. Sangeetha, "AMMU: A survey of transformer-based biomedical pretrained language models," *J. Biomed. Informat.*, vol. 126, Feb. 2022, Art. no. 103982.
[35] Y. S. Ko, J. H. Lee, and M. Song, "Examining suicide tendency social media texts by deep learning and topic modeling techniques," *J. Korean BIBLIA Soc. Library Inf. Sci.*, vol. 32, no. 3, pp. 247–264, 2021.
[36] S. Zhang, H. Yu, and G. Zhu, "An emotional classification method of Chinese short comment text based on ELECTRA," *Connection Sci.*, vol. 34, no. 1, pp. 254–273, Dec. 2022.
[37] C. Lin, P. Hu, H. Su, S. Li, J. Mei, J. Zhou, and H. Leung, "SenseMood: Depression detection on social media," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2020, pp. 407–411.
[38] D. Solse, A. Magar, P. Harde, N. Palve, and M. Jagatap, "Depression detection by analyzing social media post in machine learning using bert algorithm," *Int. Res. J. Modernization Eng. Technol. Sci.*, vol. 4, no. 4, Apr. 2022.
[39] M. El-Ramly, H. Abu-Elyazid, Y. Mo'men, G. Alshaer, N. Adib, K. A. Eldeen, and M. El-Shazly, "CairoDep: Detecting depression in Arabic posts using BERT transformers," in *Proc. 10th Int. Conf. Intell. Comput. Inf. Syst. (ICICIS)*, Dec. 2021, pp. 207–212.
[40] T. Tabak and M. Purver, "Temporal mental health dynamics on social media," 2020, *arXiv:2008.13121*.
[41] Z. Jamil, "Monitoring tweets for depression to detect at-risk users," Ph.D. dissertation, School Elect. Eng. Comput. Sci., Université d'Ottawa/Univ. Ottawa, Ottawa, ON, Canada, 2017.
[42] H. D. Zuorba, C. L. O. Olan, and A. D. Cantara, "A framework for identifying excessive sadness in students through Twitter and Facebook in the Philippines," in *Proc. Int. Conf. Bioinf. Res. Appl. (ICBRA)*, 2017, pp. 52–56.
[43] D. Mowery, C. Bryan, and M. Conway, "Feature studies to inform the classification of depressive symptoms from Twitter data for population health," 2017, *arXiv:1701.08229*.
[44] C. McClellan, M. M. Ali, R. Mutter, L. Kroutil, and J. Landwehr, "Using social media to monitor mental health discussions—Evidence from Twitter," *J. Amer. Med. Inform. Assoc.*, vol. 24, no. 3, pp. 496–502, May 2017.
[45] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of depression-related posts in reddit social media forum," *IEEE Access*, vol. 7, pp. 44883–44893, 2019.
[46] A. Roy, K. Nikolitch, R. McGinn, S. Jinah, W. Klement, and Z. A. Kaminsky, "A machine learning approach predicts future risk to suicidal ideation from social media data," *NPJ Digit. Med.*, vol. 3, no. 1, p. 78, Dec. 2020.
[47] J. Kim, Z. A. Uddin, Y. Lee, F. Nasri, H. Gill, M. Subramanieapillai, R. Lee, A. Udovica, L. Phan, L. Lui, M. Iacobucci, R. B. Mansur, J. D. Rosenblat, and R. S. McIntyre, "A systematic review of the validity of screening depression through Facebook, Twitter, Instagram, and snapchat," *J. Affect. Disorders*, vol. 286, pp. 360–369, May 2021.
[48] S. H. Pedersen, K. B. Stage, A. Bertelsen, P. Grinsted, P. Kragh-Sørensen, and T. Sørensen, "ICD-10 criteria for depression in general practice," *J. Affect. Disorders*, vol. 65, no. 2, pp. 191–194, Jul. 2001.
[49] *ICD-10 Depression Diagnostic Criteria General Practice Notebook*. Accessed: Nov. 11, 2020. [Online]. Available: https://gpnotebook.com/simplepagecfm?ID=.x20091123152205182440
[50] S. Chancellor and M. De Choudhury, "Methods in predictive techniques for mental health status on social media: A critical review," *NPJ Digit. Med.*, vol. 3, no. 1, p. 43, Dec. 2020.
[51] K. L. Clark, M. Le, Q. Manning, and C. ELECTRA, "Pre-training text encoders as discriminators rather than generators," 2020, *arXiv:2003.10555*.
[52] S. Mukherjee, A. H. Awadallah, and J. Gao, "XtremeDistilTransformers: Task transfer for task-agnostic distillation," 2021, *arXiv:2106.04563*.
[53] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," 2019, *arXiv:1909.11942*.
[54] C. Chu, T. Nakazawa, and S. Kurohashi, "Integrated parallel sentence and fragment extraction from comparable corpora: A case study on Chinese–Japanese Wikipedia," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 15, no. 2, pp. 1–22, Feb. 2016.
[55] L. N. Smith, "A disciplined approach to neural network hyper-parameters: Part 1—Learning rate, batch size, momentum, and weight decay," 2018, *arXiv:1803.09820*.
[56] P.-E. Novac, G. Boukli Hacene, A. Pegatoquet, B. Miramond, and V. Gripon, "Quantization and deployment of deep neural networks on microcontrollers," *Sensors*, vol. 21, no. 9, p. 2984, Apr. 2021.
[57] A. Rodriguez, E. Segal, E. Meiri, E. Fomenko, Y. J. Kim, H. Shen, and B. Ziv, "Lower numerical precision deep learning inference and training," *Intel White Paper*, vol. 3, pp. 1–19, Jan. 2018.
[58] P. Warden and D. Situnayake, *Tinyml: Machine Learning With Tensorflow Lite on Arduino and Ultra-Low-Power Microcontrollers*. Sebastopol, CA, USA: O'Reilly Media, 2019.
[59] Z. Liu, Y. Wang, K. Han, W. Zhang, S. Ma, and W. Gao, "Post-training quantization for vision transformer," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 28092–28103.
[60] P. Zhen, Z. Gao, T. Hou, Y. Cheng, and H.-B. Chen, "Deeply tensor compressed transformers for end-to-end object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 4, pp. 4716–4724, doi: 10.1609/aaai.v36i4.20397.

MUHAMMAD RIZWAN received the M.Sc. and M.S. (Hons.) degrees in computer science from the Islamia University of Bahawalpur, Pakistan, in 2008 and 2010, respectively, where he is currently pursuing the Ph.D. degree. He is currently working as a Full Time Lecturer with the Department of Information Technology, Khwaja Fareed University of Engineering and Information Technology. His research interests include machine learning, social informatics, data mining, and aligned areas. He received Gold Medal in M.Sc. degree.

MUHAMMAD FAHEEM MUSHTAQ received the B.S. degree in IT and the M.S. degree in CS from the Islamia University of Bahawalpur, Bahawalpur, Punjab, Pakistan, in 2011 and 2013, respectively, and the Ph.D. degree from the Faculty of Computer Science and Information Technology, University Tun Hussein Onn Malaysia (UTHM), Malaysia, in 2018. He received Microsoft certifications of Internet Security and Acceleration (ISA) Server, Microsoft Certified Professional (MCP), Microsoft Certified Technology Professional (MCTS), in 2010. He has made several contributions through research publications and book chapters toward Information Security and Artificial Intelligence. He is currently working as the Head of the Department of Artificial Intelligence, Islamia University of Bahawalpur. Previously, he was worked as the Head/Assistant Professor of the Department of Information Technology, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan, Pakistan. He was worked as a Research Assistant during the Ph.D. degree from March 2016 to August 2018. His main research interest includes information security, artificial intelligence, and cognitive system and applications.

UROOJ AKRAM received the B.S. degree in computer science from the Islamia University of Bahawalpur, Punjab, Pakistan, in 2013, and the M.S. degree in information technology from the Faculty of Computer Science and Information Technology, UTHM, Malaysia, in 2018. She is currently pursuing the Ph.D. degree in computer science with the Department of Artificial Intelligence, Islamia University of Bahawalpur. She has one year of experience as a Lecturer at the Department of Information Technology, KFUEIT, Rahim Yar Khan. She is currently working as an Associate Lecturer with the Department of Artificial Intelligence, Islamia University of Bahawalpur. Her research work is published in well-reputed high-impact journals. Her research interests include machine learning, deep learning, natural language processing, and information security.

ARIF MEHMOOD received the Ph.D. degree from the Department of Information and Communication Engineering, Yeungnam University, South Korea, in November 2017. He is currently working as an Assistant Professor with the Department of Computer Science and IT, Islamia University of Bahawalpur, Pakistan. His recent research interests include data mining, mainly working on AI and deep learning-based text mining and data science management technologies.

IMRAN ASHRAF received the M.S. degree in computer science from the Blekinge Institute of Technology, Karlskrona, Sweden, in 2010, and the Ph.D. degree in information and communication engineering from Yeungnam University, Gyeongsan, South Korea, in 2018. He has worked as a Postdoctoral Fellow at Yeungnam University. He is currently working as an Assistant Professor with the Information and Communication Engineering Department, Yeungnam University. His research interests include indoor positioning and localization, advanced location-based services in wireless communication, smart sensors (LIDAR) for smart cars, and data mining.

BENJAMÍN SAHELICES is currently working as a Professor with the Department of Informatics, University of Valladolid, Spain. His research interests include computer architecture and parallel computing.