INSTITUTE FOR HOMELAND SECURITY

Sam Houston State University

COMPARATIVE ANALYSIS OF NLP MODELS FOR DETECTING DEPRESSION ON TWITTER

Institute for Homeland Security

Sam Houston State University

Khushi Gupta,

Razaq Jinad

Qingzhong Liu

Comparative Analysis of NLP Models for Detecting Depression on Twitter

Khushi Gupta
Dept. of Computer Science
Sam Houston State University
Huntsville, Texas, USA
kxg095@shsu.edu

Razaq Jinad
Dept. of Computer Science
Sam Houston State University
Huntsville, Texas, USA
raj032@shsu.edu

Qingzhong Liu
Dept. of Computer Science
Sam Houston State University
Huntsville, Texas, USA
qxl005@shsu.edu

Abstract—Depression is a serious mental health issue affecting a significant portion of the world's population. With the widespread use of social media platforms, researchers have explored the possibility of utilizing natural language processing (NLP) techniques to detect signs of depression in users' posts. In this paper, we present a comparative analysis of six different NLP models, namely BERT, RoBERTa, DistilBERT, ALBERT, Electra, and XLNet, for depression detection on Twitter data. The experiments compare the performance of different models, and the results reveal that the highest-performing models include XLNet, DistilBERT, and RoBERTa with accuracies of over 99%.

Index Terms—Natural language processing, Depression detection, Transformers, Machine learning, Comparative analysis

I. INTRODUCTION

Depression, a common mental illness, is estimated to affect 3.8% of the worldwide population, including 5.0% among adults and 5.7% among adults older than 60 [1]. Approximately 280 million people worldwide have depression [1], which may become a severe health condition. It has different consequences for individuals and society. The economy can also be negatively affected by depression due to decreased productivity, increased absences, reduced workforce participation, and lower tax revenues. Finally, despite efforts to reduce mental health stigma, depression remains stigmatized in many societies. This can lead to discrimination, decreased access to healthcare services, and tremendous suffering for people suffering from depression.

Twitter is a popular social media platform that enables users to share short messages called tweets. It has over 330 million active users and 500 million tweets sent daily [2]. Analyzing language patterns, user behavior, and other metadata in tweets makes it possible to extract a wide range of information, from identifying consumer preferences to tracking the spread of infectious diseases. Data from Twitter can be used for various purposes [3]–[6]. Twitter is increasingly being used as a data source for political analysis. Researchers and stakeholders have analyzed tweets to track public opinion during elections, understand political communication patterns, and identify social media's role in political campaigns. Twitter has also been used to track the spread of infectious diseases such as COVID-19. Researchers can identify transmission patterns and predict epidemics by analyzing tweets related to symptoms and disease outbreaks [4]. Companies are increasingly using Twitter data as a source of customer feedback and to track consumer trends. By analyzing tweets, businesses can identify consumer preferences, monitor sentiment toward their products, and develop more effective marketing campaigns. Twitter has been used to track the response to natural disasters such as hurricanes and earthquakes. Government and emergency services can identify response patterns by analyzing tweets and helping victims.

Natural Language Processing (NLP) is a subset of artificial intelligence that encompasses how computers learn and understand human languages [7]. NLP involves various techniques and algorithms that allow

computers to process and analyze large amounts of natural language data, such as text, speech, and audio recordings. Some of the tasks that NLP can perform include Language translation sentiment analysis, speech recognition and synthesis, text classification, and topic modeling. NLP has numerous applications in various industries, such as healthcare, finance, customer service, and marketing.

Social media platforms like Twitter offer a promising new avenue for detecting depression. Twitter's vast amounts of data allow one to detect depression or its associated symptoms. Additionally, researchers have shown that analyzing language patterns and other metadata in tweets can accurately identify potential indicators of depression. In this research, we aim to use six specialized NLP models called transformers from the BERT family to analyze and detect depression from data on Twitter. The models used in the paper include BERT [8], XLNet [9], ALBERT [10], Electra [11], DistilBERT [12] and RoBERTa [13]. We aim to achieve the following goals through this project:

- We compared six transformer models to analyze and detect depression.
- We intend to improve the accuracy of detection using relatively newer transformer models.

The rest of this paper follows the following structure. In Section II, we provide details of previous research related to this topic. Then, Section III details the methodology adopted, including preprocessing and classification models. Furthermore, Section IV shows the analysis and results. Finally, Section V shows the conclusion and future works.

II. LITERATURE REVIEW

This section summarizes research studies that employ NLP and ML techniques to detect signs of depression from social media textual data.

Likewise, in the study [14], Govindasamy et al. identify depression based on Twitter content. The researchers begin by gathering tweets through Tweepy and then perform preprocessing steps, including removing URLs, retweets, mentions, and stopwords. The dataset is then tokenized, and sentiment analysis is conducted. Next, the data is input to two classifiers, Naive Bayes and NBTree (a combination of Naive Bayes and Decision Tree models). The findings indicated that both algorithms achieved almost identical accuracy levels of 97.3%, leading to equally effective models.

In [15], the authors studied health-related tweets to identify depression using machine learning algorithms like Multinomial Naive Bayes and Support Vector Regression. They first preprocessed the data and then extracted features using techniques like stemming, tokenization, gram features, POS vectorization, and sentiment analysis. After that, they fed the data into various classifiers, including Naive Bayes, K-means clustering, and SVM. The results showed that the Support Vector Regression classifier proposed by the authors had the highest accuracy of 79.7% among all the tested algorithms.

Furthermore, [16] present a system that identifies users at risk of depression from their Twitter posts. Toward this, the authors present an efficient neural network architecture that improves and optimizes word embeddings. They evaluate the optimized embeddings produced by the architecture along with three commonly used word embeddings, random trainable, skip-gram, and CBOW, on the CLPsych 2015 shared task and the Bell Let's Talk datasets. They then compare the performance of CNN-based and RNN-based models to determine the best models and parameters for depression detection. The experiments reveal that CNN-based models perform better than RNN-based models.

In [17], the authors suggest a combined CNN and LSTM model for identifying people with depression using conversational text data obtained from Twitter. They apply their proposed model and machine-learning

classifiers on a Twitter dataset to compare their performance in detecting depression. The authors report that the proposed model yields an accuracy of 92%, which outperforms the maximum accuracy of 83% achieved by the machine learning technique.

Additionally, The article [18] presents an approach that combines natural language processing (NLP) techniques and machine learning (ML) algorithms to classify textual data. The authors optimized the pretrained RoBERTa model using several techniques, such as adding a normalization layer, a Dropout layer to prevent overfitting and a linear layer for classification. By implementing these modifications, the authors achieved a higher accuracy of 96% compared to other models in the BERT family, including BERT, DistilBERT, MP Net, RoBERTa, and ALBERT, which reached a maximum of 92%.

Upon studying the existing work in the literature, to the best of our knowledge, no work has tackled the usage and comparison of different BERT family natural language processing models to detect depression from Twitter data.

III. METHODOLOGY

In this research, we test various model variants of BERT to determine the most accurate model for classifying depressive tweets. Figure 1 depicts the methodology used for this research.

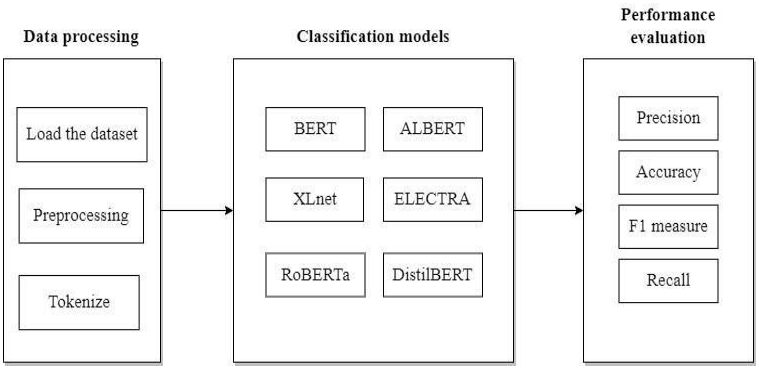

Fig. 1. Proposed Framework

A. Dataset

The dataset used for this research consists of two data sources. We merged two different datasets to get the final dataset for our research. In order to differentiate depressive tweets online, we used the “Depression detection using Twitter post” dataset from Github (https://github.com/eddieir/Depression detection using Twitt er post). This dataset contains depressive tweets collected using the TWINT web scraper using the keyword” depression.” This dataset was then appended to the “Twitter Sentiment Analysis” dataset from Kaggle (https://www.kaggle.com/code/sharanharsoor/twittersentiment-analysis/data). This dataset contains tweets based on three categories negative, neutral, and positive. For our dataset, we extracted all the positive tweets to add to the dataset with the depressive tweets. Since the data of positive tweets extracted was a lot compared to the depressive tweets, we sampled the data by randomly selecting 3500 rows from the data. This was done to avoid class imbalance. The two different data were then merged to create one dataset of depressive and positive tweets of 5845 rows and two columns, namely text and target (where 0 was a depressive tweet whereas 1 was a positive tweet).

B. Preprocessing

The methodology of our research starts with preprocessing of the dataset. The dataset contains raw data in the form of sentences with special symbols, usernames, hashtags, and URLs. Such types of data are not required for machine learning models. Thus, preprocessing is needed, which deals with data preparation and transformation of the dataset to make the input data easier to decode and interpret.

Initially, all the characters are converted to lowercase. After that, all the special characters such as emojis, html code, URLs, Unicode characters, digits, mentions, and punctuation are eliminated. This can reduce noise and improve the accuracy of the model.

We then remove stop words such as "the," "a," "an," and "in." Stopwords are words that do not add meaningful content to the dataset (i.e., pronouns, prepositions, conjunctions, etc.) Thus, removing stopwords can reduce the dimensionality of the data and improve the model's accuracy. This is necessary for the simplification of the language and efficient vectorization.

Lastly, Lemmatization is performed using WordNetLemmatizer. Lemmatization reduces the inflectional forms of words to their base or root form. This can reduce the number of unique words in the data and make it easier for the machine-learning algorithm to understand the underlying meaning.

C. Classification models

A transformer is a neural network that transduces input data into vector representations using self-attention layers with encoder–decoder structure [19]. In this study, we compared different natural language processing models from the BERT family and performed a comparative study on the results produced by the different models in terms of accuracy in the classification of depression in Twitter posts.

1) BERT: BERT, short for Bidirectional Encoder Representations from Transformers, is a powerful Natural Language Processing (NLP) model that was introduced by Google in 2018. It is a pre-trained language model that can understand the context of words in a sentence, allowing it to generate high-quality text representations. What sets BERT apart from other NLP models is its ability to process bidirectional text sequences, meaning it can take into account the context of the words both before and after the current word. This makes BERT extremely effective for a wide range of NLP tasks.

2) ALBERT: ALBERT (A Lite BERT) is a language model developed by Google Research as an improvement over BERT. ALBERT addresses the problem of parameter redundancy in BERT, which led to high memory consumption and slow training times. ALBERT achieves this by utilizing parameter sharing across layers, groups of layers, and across both the encoder and decoder in a transformer model. Additionally, ALBERT introduces a self-supervised loss based on sentence order prediction, which improves performance on downstream NLP tasks.

3) XLNet: XLNet is a state-of-the-art language model that introduces new ideas, such as a permutationbased approach to pretraining that enables the model to consider all possible orderings of the input sequence and a new objective function that maximizes the expected likelihood of generating each word in the input sequence conditioned on all possible permutations of the other words. This enables XLNet to capture bidirectional dependencies without masking tokens during pretraining.

4) ELECTRA: ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) is a language model that introduces a new pretraining method for language models. ELECTRA trains a discriminator network to distinguish between real and fake tokens generated by a generator network. This allows ELECTRA to use all tokens during pretraining rather than masking some, which makes it more efficient. The generator network is trained to replace some tokens in the input sequence with generated

tokens, while the discriminator network is trained to predict whether each token in the sequence is real or generated.

5) RoBERTa: ROBERTA (Robustly Optimized BERT Pretraining Approach) improves on BERT by using a larger training corpus and training for longer, which allows it to learn more nuanced language features. ROBERTA uses dynamic masking during pretraining, which involves randomly masking different spans of text in each training instance rather than always masking the same tokens. This forces the model to learn more about the context of words and phrases.

6) DistilBERT: DistilBERT is a smaller and faster version of BERT. It achieves this by using a distillation technique that compresses the original BERT model while retaining most of its performance. Specifically, DistilBERT removes some of the unnecessary parameters from BERT and uses a smaller transformer architecture. Despite its smaller size, DistilBERT achieves comparable performance to the original BERT model on several NLP tasks.

IV. RESULTS

For this research, we analyze the results in the form of accuracy, precision, recall, and f1-score. Classification accuracy is the percentage of correctly predicted incidents to the total incidents.

$$\frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

Where:

TP (True Positive) = Number of correctly predicted depressive tweets TN

(True Negative) = Number of correctly predicted nondepressive tweets FP

(False Positive) = Number of incorrectly predicted depressive tweets

FN (False Negative) = Number of incorrectly predicted nondepressive tweets

Using classification accuracy to assess model performance alone is not ideal. Therefore, additional metrics should be used to assess each model's performance on the test set for the classification task. Precision, recall, and F1 scores generated by the Python package Scikit-learn [20] were used to achieve a fair model assessment.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1Score = (\frac{Recall * Precision}{Recall + Precision}) \quad (4)$$

Precision measures the correctly classified positive cases from all the predicted positive cases. Recall is used to quantify the number of correctly classified positive cases made out of all positive cases in the dataset. F1 score is the harmonic mean of Precision and Recall; it combines both measurements into a single measure. We then used the macro average to evaluate the overall performance of each model. Table 1 depicts the result measures for all the utilized models.

A. BERT

To preprocess our data using BERT, we utilized the "bert en uncased preprocess version 3" from Tensorflow. This model gets its vocabulary for English from Wikipedia and BooksCorpus. It implements the preprocessor API that normalizes the text inputs by converting them into lowercase before tokenization into words and removes any accent markers. The output of this model is a series of fixed-length sentences that can be input into the transformer model.

For classification, we used the "bert en uncased L-12 H768 A-12" model from Tensorflow. It uses L=12 hidden layers, a hidden size of H=768, and A=12 attention heads. It implements the encoder API for embedding text. We trained this model with 15 epochs and a batch size of 32, which yielded a test accuracy of 90.19%. Figures 2 and 3 exhibit the train and validation accuracy and loss plots for BERT.

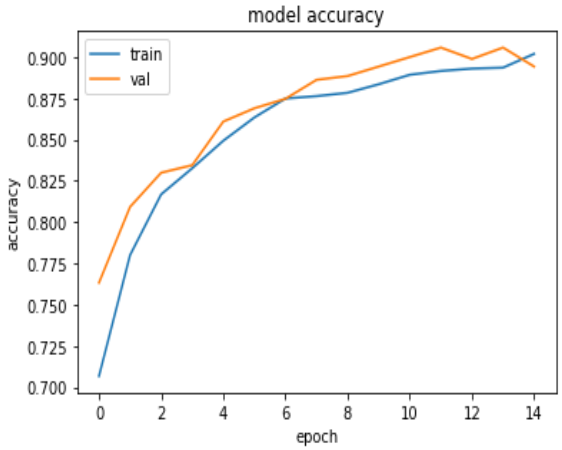

Fig. 2. Accuracy plot for BERT

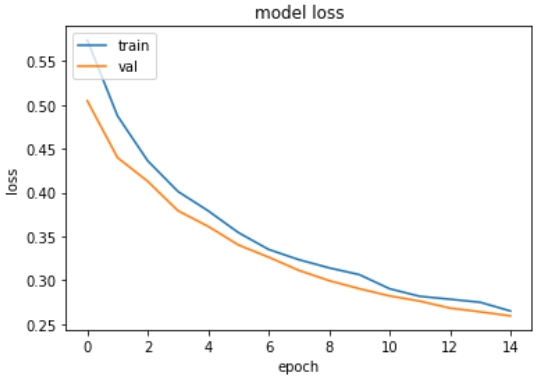

Fig. 3. Loss plot for BERT

B. ALBERT

We utilized "albert en preprocess version 3" to train our data using ALBERT. Its functionality is similar to the preprocessing model utilized in BERT. For classification, we used the "albert en base 3" model. We trained this model with 12 epochs and a batch size of 32, which yielded a test accuracy of 84.79%. Figures 4 and 5 show the train and validation accuracy and loss plots for ALBERT.

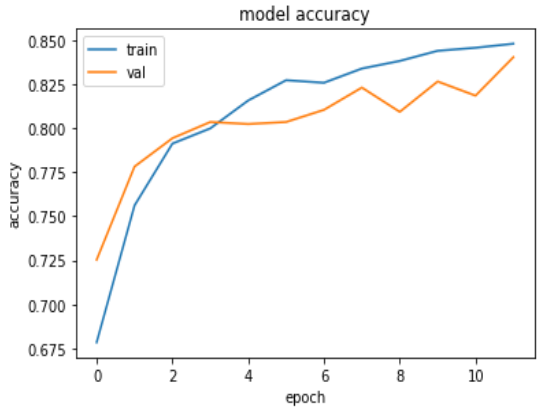

Fig. 4. Accuracy plot for ALBERT

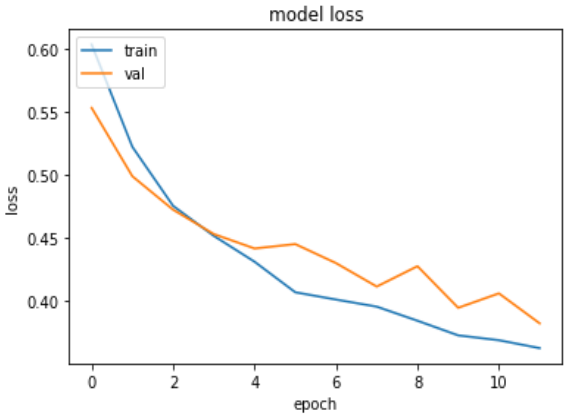

Fig. 5. Loss plot for ALBERT

C. DistilBERT

For DistilBERT classification, we used the "distilbert-baseuncased-finetuned-sst-2-english" from Hugging face. It uses the Stanford Sentiment Treebank(sst2) corpora for the model. We trained this model with 3 epochs with a batch size of 16, yielding a test accuracy of 99.84%. Figures 6 and 7 depict the train and validation accuracy and loss plots for DistilBERT.

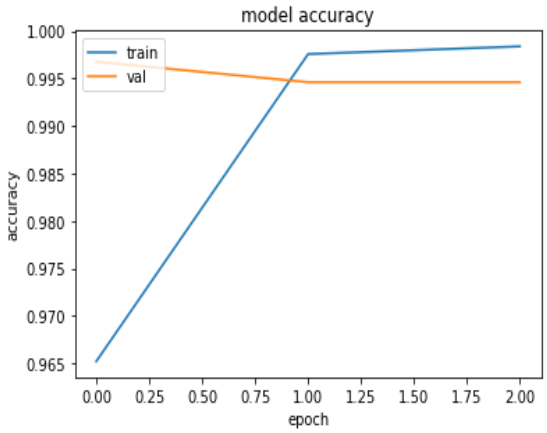

Fig. 6. Accuracy plot for DistilBERT

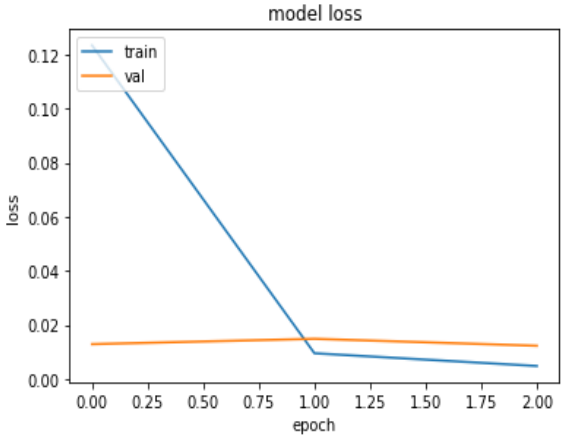

Fig. 7. Loss plot for DistilBERT

D. ELECTRA

To train our data using ELECTRA, we utilized the same preprocessing model as BERT, which is “bert en uncased preprocess version 3” from Tensorflow. For classification, we used the “electra base version 2” model from Tensorflow. We trained this model with 10 epochs with a batch size of 32, yielding a test accuracy of 87.09%. Figures 8 and 9 depict the train and validation accuracy and loss plots for Electra.

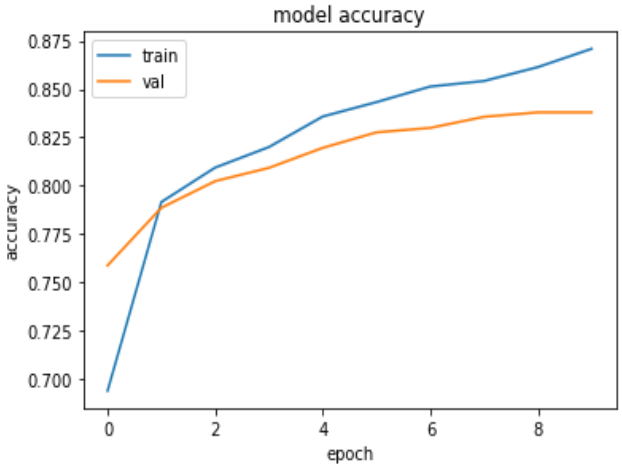

Fig. 8. Accuracy plot for ELECTRA

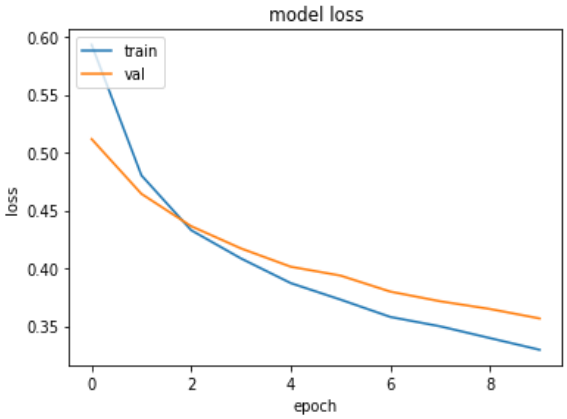

Fig. 9. Loss plot for ELECTRA

E. RoBERTa

For RoBERTa classification, we used the "Roberta-base" from Hugging face. This model utilizes masked language modeling (MLM). The process involves randomly masking 15% of the words in a sentence, then feeding the masked sentence into the model to predict the missing words. We trained this model with 3 epochs with a batch size of 8 and a learning rate of 1e-5. This yielded us a test accuracy of 99.9%. Figures 10 and 11 display the train and validation accuracy and loss plots for RoBERTa.

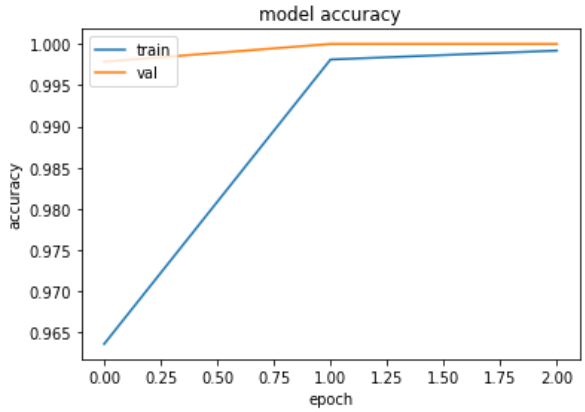

Fig. 10. Accuracy plot for RoBERTa

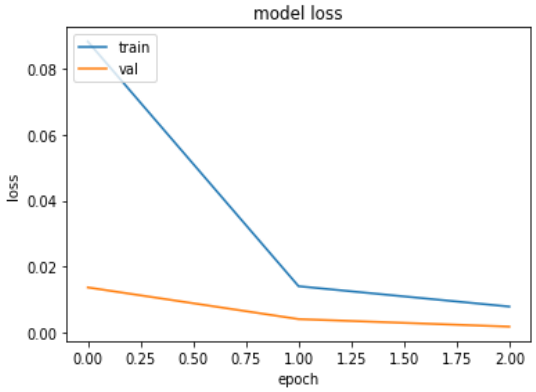

Fig. 11. Loss plot for RoBERTa *F.*

XLNet

For XLNet classification, we used the "xlnet-base-cased" from Hugging face. It is an approach to unsupervised language representation learning, which utilizes an innovative generalized permutation language modeling objective. Moreover, XLNet utilizes Transformer-XL as its underlying model architecture, which performs exceptionally well for languagerelated tasks that require the processing of lengthy contexts. We trained this model with 5 epochs with a batch size of 32, yielding a test accuracy of 99.9%. Figures 12 and 13 display the train and validation accuracy and loss plots for XLNet.

V. CONCLUSION AND FUTURE WORK

In this paper, we presented a comparative analysis of various natural language processing models for detecting depression on Twitter. We explored six different models: BERT, RoBERTa, DistilBERT, ALBERT, Electra, and XLNet. Our experiments show that RoBERTa, XLNet, and DistilBERT outperformed all the other models with an accuracy of 99%. Our study thus highlights the importance of selecting an appropriate model for depression detection on Twitter. In the future, we aim to explore the use of multi-modal data, such as text and pictures, in conjunction with the BERT family NLP models to improve the accuracy of depression

detection using Twitter data. Moreover, investigating the generalization of the models on datasets collected from other social media platforms can also be a fruitful direction for future research.

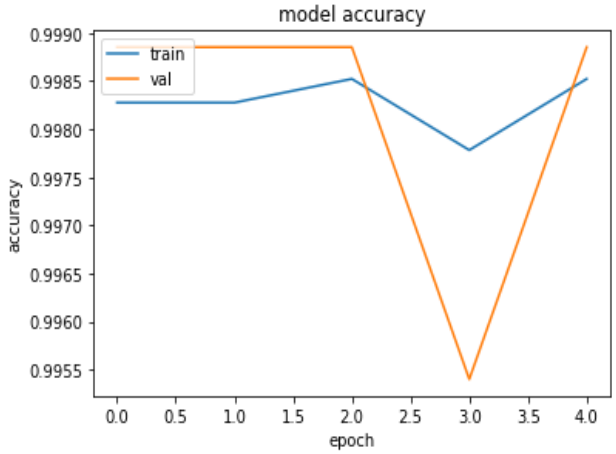

Fig. 12. Accuracy plot for XLNet

TABLE I TRAINING AND TESTING METRICS FOR THE NLP MODELS

	Training Metrics		Testing Metrics			
	Accuracy	Loss	Accuracy	Precision	Recall	F-Measure
XLNet	0.1000	0.0000	0.9954	1.00	0.99	0.99
DistilBERT	0.9984	0.0048	0.9982	0.4101	1.0	0.5663
ALBERT	0.8479	0.3621	0.8503	0.8506	0.8344	0.8405
BERT	0.9019	0.2648	0.8979	0.9146	0.8772	0.8892
Electra	0.8709	0.3296	0.8565	0.8505	0.8501	0.8503
RoBERTa	0.999	0.008	0.999	1.0	0.9978	0.9989

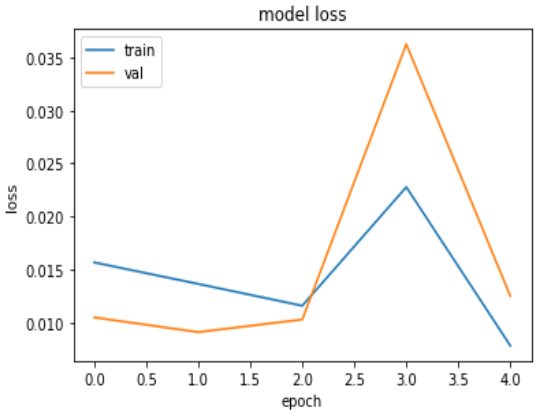

Fig. 13. Loss plot for XLNet

ACKNOWLEDGMENT

The support for this study from the SHSU Institute of Homeland Security is highly appreciated.

REFERENCES

[1] "Global burden of disease study 2019 (gbd 2019) data resources." [Online]. Available: https://ghdx.healthdata.org/gbd-2019

[2] W. T. Matt Ahlgren, "Top 55+ twitter statistics, facts amp; user demographics for 2023," Mar 2023. [Online]. Available: https://www.websitehostingrating.com/twitter-statistics/

[3] A. Ceron, L. Curini, and S. M. Iacus, *Politics and big data: Nowcasting and forecasting elections with social media*. Taylor & Francis, 2016.

[4] P. Grover, A. K. Kar, and G. Davies, ""technology enabled health"– insights from twitter analytics with a socio-technical perspective," *International Journal of Information Management*, vol. 43, pp. 85–97, 2018.

[5] A. Karami, A. A. Dahl, G. Turner-McGrievy, H. Kharrazi, and G. Shaw Jr, "Characterizing diabetes, diet, exercise, and obesity comments on twitter," *International Journal of Information Management*, vol. 38, no. 1, pp. 1–6, 2018.

[6] T. D. Nascimento, M. F. DosSantos, T. Danciu, M. DeBoer, H. van Holsbeeck, S. R. Lucas, C. Aiello, L. Khatib, M. A. Bender, U. U. G. C. O. 2014 *et al.*, "Real-time sharing and expression of migraine headache suffering on twitter: a cross-sectional infodemiology study," *Journal of medical Internet research*, vol. 16, no. 4, p. e96, 2014.

[7] A. Chopra, A. Prashar, and C. Sain, "Natural language processing," *International journal of technology enhancements and emerging engineering research*, vol. 1, no. 4, pp. 131–134, 2013.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[9] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[10] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for selfsupervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[11] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pretraining text encoders as discriminators rather than generators," *arXiv preprint arXiv:2003.10555*, 2020.

[12] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[14] K. A. Govindasamy and N. Palanichamy, "Depression detection using machine learning techniques on twitter data," in *2021 5th international conference on intelligent computing and control systems (ICICCS)*. IEEE, 2021, pp. 960–966.

[15] P. Arora and P. Arora, "Mining twitter data for depression detection," in *2019 international conference on signal processing and communication (ICSC)*. IEEE, 2019, pp. 186–189.

[16] A. H. Orabi, P. Buddhitha, M. H. Orabi, and D. Inkpen, "Deep learning for depression detection of twitter users," in *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*, 2018, pp. 88–97.

[17] B. Verma, S. Gupta, and L. Goel, "A neural network based hybrid model for depression detection in twitter," in *Advances in Computing and Data Sciences: 4th International Conference, ICACDS 2020, Valletta, Malta, April 24–25, 2020, Revised Selected Papers 4*. Springer, 2020, pp. 164–175.

[18] B. Rodrawangpai and W. Daungjaiboon, "Improving text classification with transformers and layer normalization," *Machine Learning with Applications*, vol. 10, p. 100403, 2022.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[20] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler *et al.*, "Api design for machine learning software: experiences from the scikit-learn project," *arXiv preprint arXiv:1309.0238*, 2013.

INSTITUTE FOR HOMELAND SECURITY

Sam Houston State University

The Institute for Homeland Security at Sam Houston State University is focused on building strategic partnerships between public and private organizations through education and applied research ventures in the critical infrastructure sectors of Transportation, Energy, Chemical, Healthcare, and Public Health.

The Institute is a center for strategic thought with the goal of contributing to the security, resilience, and business continuity of these sectors from a Texas Homeland Security perspective. This is accomplished by facilitating collaboration activities, offering education programs, and conducting research to enhance the skills of practitioners specific to natural and human caused Homeland Security events.

Institute for Homeland Security
Sam Houston State University

Gupta, K., Jinad, R., & Liu, Q. (2023) Comparative Analysis of NLP Model for Detecting Depression on Twitter. (Report No. IHS/CR-2023-1011). The Sam Houston State University Institute for Homeland Security. https://doi.org/10.17605/OSF.IO/J5QF3