A Multilevel Depression Detection from Twitter using Fine-Tuned RoBERTa

Afra Zaman
Department of ICT
Bangladesh University of Professionals
Dhaka, Bangladesh
propazaman12@gmail.com

Syeda Sunjida Ferdous
Department of ICT
Bangladesh University of Professionals
Dhaka, Bangladesh
sunjidaferdous123@gmail.com

Nazneen Akhter
Department of CSE
Bangladesh University of Professionals
Dhaka, Bangladesh
nazneen.akhter@bup.edu.bd

Tabassum Ibnat Ena
Department of ICT
Bangladesh University of Professionals
Dhaka, Bangladesh
tabassumibnat245@gmail.com

Md. Mahmudun Nabi
Department of ICT
Bangladesh University of Professionals
Dhaka, Bangladesh
mahmudshafim@gmail.com

Salma Akter Asma
Department of ICT
Bangladesh University of Professionals
Dhaka, Bangladesh
aaliyaasma786@gmail.com

***Abstract*—Depression stands out as the most commonly occurring psychiatric disorder worldwide. Extreme levels of depression can make people take drastic steps. However, early-stage depression detection and treatment can save many lives. Different social media platforms play a vital role in analyzing depression in its early stages. Hence, this work aims to identify and intensify depression through social media posts. This research highlights the potential of social media for the early detection of depression and its significance in aiding psychiatrists to prevent suicide by identifying depression in its initial phases. The study employs a fine-tuned RoBERTa model to forecast multilevel depression (mild, moderate, and severe) from Twitter posts. The research aim to enhance comprehension of depression severity and its impact on mental health by conducting a thorough context analysis utilizing a fine-tuned RoBERTa model. The model proposed in this study exhibits a 90% of accuracy in predicting multilevel depression. The analysis illuminates the effectiveness of the suggested framework and its possible implementation in mental health research and treatment.**

***Keywords*—** Multilevel Depression, Fine-tuned RoBERTa, Context Analysis, Social Media, Early Detection

I. Introduction

Depression is a common mental health problem that makes people feel sad, empty, and uninterested in things they used to enjoy. It has a potential effect on an individual's interpersonal relationships, professional or academic achievements, physical well-being, and overall quality of life. The present study highlights the association between depression and increased susceptibility to various medical conditions, including cardiovascular diseases, as well as elevated mortality rates in the population [1]. The prevalence of geriatric depression in China, as reported by Suhara et al. has been found to be concerning, with rates ranging from 11.5% to 21.1% [2].

Over the past decade, social media has become a platform where a people share their everyday activities, experiences and thoughts. This platform provides information to identify mental health issues. [3]. The prevention and early detection of mental illness have been a topic of interest among researchers, who have sought to encourage these efforts [4]. Mustafa et al. have presented a comparison work among NN, SVM, RF, and 1DCNN where they only focused on the word level attention [4]. Moreover, researchers only limited their works to binary classification (depressed and non-depressed) [5], [6].

This study tries to fill the gap between those in need of assistance and the professionals who can offer it. The goal of this study is to employ a different methodology for the early identification of depressed people. In recent research trends, researchers are focusing on finding out only about depressed and not-depressed people. They are not giving enough highlights to the level of depression. It is essential to find out the level of depression so that the doctors can get an idea of the patient's depressive stage. Moreover, sentence-level context analysis can enhance the performance of the depression detection model more than word-level context analysis.

For this investigation, Twitter posts are considered that have been divided into three categories: mild, moderate, and severe, based on the level of depression. The model utilized in this study is based on RoBERTa, which has a better pre-training process and is similar to BERT architecture. The goal of this study is to use RoBERTa's strong ability to classify unstructured data utilizing sentence-level context analysis, especially user-generated texts, to classify different kinds of mental illness. This study wants to find out if RoBERTa can correctly classify online text about mental health conditions by using its good performance and ability to learn contextual knowledge.

II. Literature Review

In the year 2011, Moreno et al. used a Machine Learning (ML) based protocol to detect depression in college students using Facebook posts, identifying depression signs in 25% of profiles and major depressive episodes in 2.5%, highlighting the potential of social media for detection [7].

Shen et al. developed a machine-learning model to identify depression using Twitter. They proposed two dictionaries namely Multimodal Depressive Dictionary Learning (MDL) and Wasserstein Dictionary Learning (WDL). The Naive Bayesian classifier with MDL performs at an 85% F1 score. The model works best for detecting moderate depression (40%-60%) [6]. Different data mining techniques on Reddit have been applied to identify suicide discussions, aiming to enhance analytical methods and collaborate for broader scope [8]. Hasan et al. has employed ML and sentiment analysis to measure depression, achieving high accuracies of 91% (SVM), 83% (NB), and 80% (ME) [9]. Zucco et al.and Birjalia et al. have utilized sentiment analysis to predict depression and suicidal thoughts from social media posts [10], [11]. Almeida et al. employ an ensemble classification approach to predict depression risk by analyzing user-generated content on Reddit [12]. Alhanai et al. used LSTM to identify depression risk through audio and text achieving a 0.77 F1 score [13]. Madhu has used concepts like Bag of Words, Part of Speech, and Natural Language Processing to detect potential signs of suicide risk by analyzing sentiment and content in online text [14]. Coppersmith et al. examine suicidal risk in social media data by utilizing BiLSTM model [15]. Stephen et al. utilize lexicon to develop an accurate model for early depression detection on social media [16]. Burdissoa et al. proposed an explainable SS3 model to automate depression risk recognition in Twitter, achieving an F1 score 0.61 [17]. Mustafa et al.used NN, SVM, RF, and 1DCNN to detect depression on Twitter. They analyzed tweets from 179 individuals in depression treatment, applying pre-processing techniques and TF-IDF weighting. The model determines depression levels as high, mid, and low. 1DCNN achieved 91% accuracy.They considered word-level context analysis [4]. Murarka et al. proposed a DL method using a RoBERTa-based classifier to identify and categorize five prevalent mental diseases by analyzing unstructured user data from social media networks. The study utilized emotional discussions from a special forum on Reddit, resulting in a large and high-quality dataset. The RoBERTa-based classifier outperformed the baseline LSTM model and achieved an accuracy of 89%, comparable to BERT. The authors also discussed the potential for creating a multi-label classifier to address individuals experiencing more than one mental disorder simultaneously [3]. Logistic Regression achieved 90% of accuracy to detect depression from Bengali social media posts followed by a proposed Bengali dictionary [18]. COVID time depression has been addressed by Sharma et al. by proposing COVID-19 Emotion Analyzer [19]. Baek et al.proposed a prediction system based on Context-DNN with an accuracy of 0.95 [20].

Lin et al. developed SenseMood, a multimodal system that combines textual and visual features from Twitter data for depression detection. The model has 88.4% accuracy [5]. A RoBERTa-base framework TWEETVAL is proposed by Barbieri et al. [21]. STATENet model achieved 85.1% accuracy to identify suicidal tendencies in social media posts [22]. In the year 2021, Govindasamy et al. achieved a high accuracy of 97.31% in detecting depression using NB and NBTree classifiers on Twitter data [23]. Li et al. developed an ML model using Multi-Task Learning (MTL) to enhance depression detection in clinical interviews, achieving a 74.5% accuracy [24]. Ghosal & Jain utilized XGBoost Classifier for detecting depression from Reddit dataset with an accuracy of 71.05% [25]. Bucur et al. (2023) propose a time-enriched multimodal transformer architecture for depression identification in social media, outperforming existing approaches on a Twitter dataset [26]. Researchers are giving less focus to the multilevel prediction of depression from a Twitter post. This level of depression can help the psychiatrist to give faster treatment to their patients. Moreover, sentence-level context analysis can enhance the accuracy of the prediction model.

III. Multilevel Depression Detection Model

RoBERTa, Facebook AI's language model, is extensively trained using 1 million steps and 8,192 tokens in English Wikipedia and BookCorpus data. This comprehensive training helps RoBERTa understand complicated circumstances and deliver better contextual representations. It uses the self-attention mechanism to let the model record word connections in a phrase or sequence by focusing on distinct input portions [27]. This attention mechanism helps the model to find the exact context of the Twitter posts.

In this work, a fine-tuned Robustly Optimized BERT Pre-training Approach (RoBERTa) model is used to predict multilevel depression (mild, moderate, and severe) from Twitter posts. Incorporating two more dense layers and modifications to the underlying RoBERTa architecture developed the proposed model to learn higher-level features and patterns specific to the classification task. The working procedure of the proposed model is discussed in this section.

In the input layer, the encoded texts are tokenized using the RoBERTa tokenizer. Subsequently, the RoBERTa base model is initialized with the pre-trained weights of "Roberta-base". The added features to the pre-trained model are two dropout layers and two dense layers, and the last two frozen layers of Roberta are unfrozen.

This model's outputs go through two dropout layers. Dropout is a regularization method that involves randomly setting a portion of input units to zero during training in order to prevent overfitting. Following the dropout layers, two dense layers are incorporated. In the initial dense layer, there are 128 units utilizing ReLU activation and the L1 regularization function. This L1 regularization function helps reduce the overfitting problem. The second dense layer is comprised of units that are equivalent in number to the output labels and implements softmax activation for the purpose of multi-class classification. In the fine-tuning process, the last two layers of the RoBERTa model are unfrozen and trained for the particular text classification task.

The dataset is divided into three subsets, namely train, test, and validation sets, to aid in model training and evaluation. The model is compiled using the Adam optimizer, which adjusts the learning rate to 2e-5 throughout training. Also, use

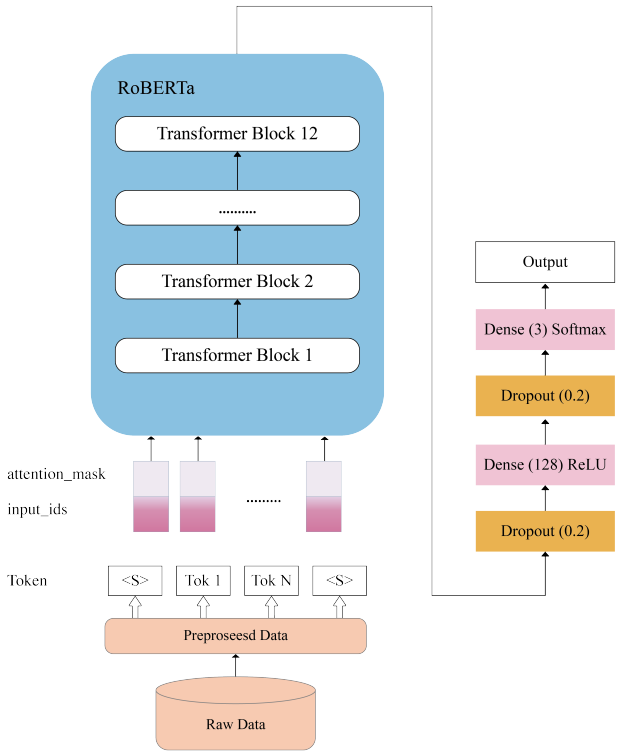

Fig. 1. Proposed fine-tuned RoBERTa model for multilevel depression detection

the sparse categorical cross-entropy loss as the loss function. The model has gone through 15 epochs and has been trained using 16 batch sizes. The evaluation of the trained model's performance is conducted on a distinct test dataset utilizing loss and accuracy metrics. Figure 1 represents the architecture of the proposed model.

A. Dataset

The study utilized a pre-existing Twitter dataset provided by Shen et al. (2017) [6]. The dataset has been classified into three distinct sections, specifically labeled D1, D2, and D3. The datasets D1 and D2 are labeled as depression and nondepression, respectively. The dataset for D3 comprises 36,993 potential users experiencing depression, and it remains unlabeled. The tweet subsets D1, D2, and D3 contain 6493, 5384, and 58900 tweets, respectively. Only English-language tweets are considered for analysis, despite the presence of tweets in various other languages. The study focuses on multiclass depression detection, and thus, the nondepressive dataset D2 is excluded. Only the D1 and D3 datasets are selected for the experiment.

1) Data annotation: In this study, 6470 tweets are extracted from the D1 dataset and subsequently classified into three levels of depression: "mild", "moderate", and "severe". The classification is performed using a transformer-based model called "cardiffnlp/twitter-roberta-base-sentiment" [28] from Hugging Face. This model is a variant of the RoBERTa model that has been specifically trained for sentiment analysis on Twitter data. Three thresholds are established in the "negative" column to determine the degree of depression present in the tweets. The following thresholds are considered using the concept of Mustafa et al. [4].

- If the value of 'negative' is greater than or equal to 80, the value of 'level' is assigned as severe.
- If the value of 'negative' is less than 80 and greater than or equal to 30, the value of 'level' is classified as moderate.
- If the value of 'negative' is less than 30, the corresponding value of 'level' is set to mild.

And then, in the case of the D3 dataset, first of all, we extracted 18819 tweets and then used the same model for labeling. After that, we followed exactly the same approach and categorized the depression tweets into three levels. Finally, The annotated dataset has been verified by some final-year psychology students.

2) Data Pre-processing: Various pre-processing techniques are employed to eliminate redundancy and noise from the dataset. The process involves eliminating URLs, non-ASCII characters, and disallowed characters, and converting emojis to text. Subsequently, the cleaned text undergoes tokenization utilizing the RoBERTa tokenizer. The tokens undergo processing to eliminate the presence of a leading character 'Ġ', if any. This procedure eliminates the additional space. Ultimately,

the tokens are reassembled into a unified string, with space between tokens separating each individual token.

The distribution of mild, moderate, and severe cases is 13.4%, 47.5%, and 39.1%, respectively. Upsampling technique has been used to remove the unbalanced relationship. This technique refers of duplicating instances in the minority class in a random manner to achieve a balanced class distribution.

IV. Performance Analysis of Fine-tuned RoBERTa

In this section, the outcomes of the proposed Fine-tuned RoBERTa model for predicting multilevel depression detection are presented. This model is fine-tuned using multiple depression labels, specifically mild, moderate, and severe, to gain an understanding of the severity of depression conveyed in Twitter posts. The model endeavors to identify indicators of extreme suicidal thoughts by analyzing the text content, language patterns, and context within these posts.

The experiment is performed using a MacBook M1 Air equipped with 8 GB of RAM and macOS Ventura 13.3.1(a). Google Colab with GPU support is utilized to create and train the proposed model in an efficient and capable working environment. The model utilized in this study integrates the RoBERTa base model along with supplementary layers, including dropout layers and dense layers, in order to enhance its capabilities in text classification. Here, four performance metrics, namely precision, recall, F1 score, and accuracy, have been used. The detailed analysis of the RoBERTa model is discussed in the following section.

The training process has been done using a batch size of 16 and a maximum of 15 epochs. The monitoring of the training process is conducted through the utilization of the validation dataset, and the restoration of the optimal weights is determined by the validation loss. Following the training phase, the model underwent evaluation on both the test and validation datasets in order to gauge its performance.

TABLE I
Test Dataset Classification Report

Category	Precision	Recall	F1 Score	Support
Mild	0.97	0.96	0.97	1224
Moderate	0.84	0.87	0.85	1224
Severe	0.89	0.87	0.88	1224
Accuracy	0.9			

Table I shows the testing performance of the model. The model achieved an accuracy rate of 90%, indicating that it effectively classified the majority of instances. The achieved precision for all classes are high, with values of 0.97, 0.84, and 0.89. The findings indicate that, the proposed model perform well in terms of accuracy utilizing two adding layers and sentence level context analysis. Figure 2 represents the precision, recall, F1 score, and accuracy for the test dataset.

Table II shows the validation dataset's classification report, which verifies the RoBERTa model's excellent performance in predicting depressive level from tweets. The model successfully detected examples from each class with accuracies of 0.97, 0.85, and 0.89 for the mild, moderate, and severe

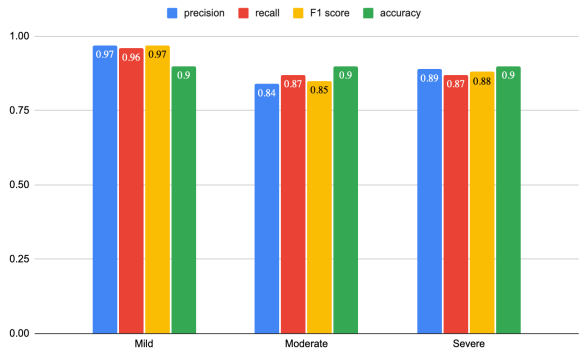

Fig. 2. Performance on the test dataset

TABLE II
Validation Dataset Classification Report

Category	Precision	Recall	F1 Score	Support
Mild	0.97	0.96	0.96	1224
Moderate	0.85	0.86	0.86	1224
Severe	0.89	0.89	0.89	1224
Accuracy	0.9			

categories, respectively. Both the recall and F1 scores are in the 0.86 to 0.96 range, showing that the model is able to accurately identify true positives and produce consistent results. The results of this study demonstrate the promise of this approach for both applied settings and future investigations into the detection and analysis of depression. Precision, recall, F1 score, and accuracy are depicted in a bar chart in Figure 3.

The accuracy scores achieved by different existing studies in classifying depression using various datasets and models are summarized in Table III. The proposed model fine-tuned RoBERTa attained a peak accuracy of 90% for Twitter data. The existing RoBERTa model by Murarka et al. achieved 89% accuracy [3]. In a separate study, the STATENet model attained an accuracy of 85.1% [22]. The accuracy of alternative models, including XGBoost and MTL, ranged from 71.05% to 74.5% [25], [24]. The presented performance demonstrates the potential usefulness of the model in diverse applications and

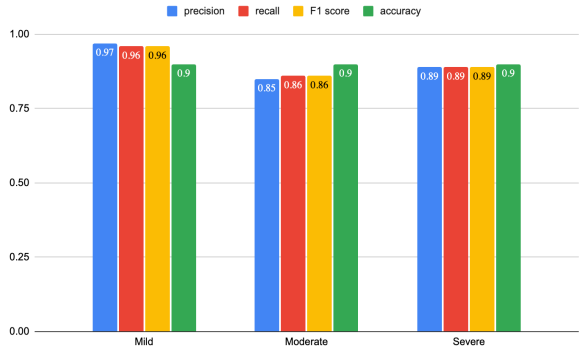

Fig. 3. Performance on the validation dataset

TABLE III
COMPARISON OF DEPRESSION DETECTION MODELS

Authors	Dataset source	Used model	Accuracy
Ghosal and Jain 2023 [25]	Reddit	XGBoost	71.05
Li et al. 2022 [24]	Clinical interviews	MTL	74.5
Shen et al. 2017 [6]	Twitter (Pu)	MDL	84.8
Sawhney et al. 2020 [22]	Twitter (Pu)	STATENet	85.1
Murarka et al. 2020 [3]	Reddit	RoBERTa	89
Proposed model	Twitter (Pu)	fine-tuned RoBERTa	90

research studies concerning depression detection and analysis.

CONCLUSION

In this study, Twitter posts has been considered for multilevel (mild, moderate, severe) depression detection utilizing fine-tuned RoBERTa. The multilevel detection model outperforms with 90% accuracy than the base model considering sentence level context analysis. Understanding the severity of depression can help healthcare providers better assess patients' psychological well-being and identify those at risk of severe suicidal thoughts. In future, the annotation will be validated with the help of a professional psychiatrist.

REFERENCES

[1] H. K. Koh and A. K. Parekh, "Toward a united states of health: implications of understanding the us burden of disease," *JAMA*, vol. 319, no. 14, pp. 1438–1440, 2018.

[2] S. Lin, Y. Wu, L. He, and Y. Fang, "Prediction of depressive symptoms onset and long-term trajectories in home-based older adults using machine learning techniques," *Aging & Mental Health*, pp. 1–10, 2022.

[3] A. Murarka, B. Radhakrishnan, and S. Ravichandran, "Detection and classification of mental illnesses on social media using roberta," *arXiv preprint arXiv:2011.11226*, 2020.

[4] R. U. Mustafa, N. Ashraf, F. S. Ahmed, J. Ferzund, B. Shahzad, and A. Gelbukh, "A multiclass depression detection in social media based on sentiment analysis," in *17th International Conference on Information Technology–New Generations (ITNG 2020)*. Springer, 2020, pp. 659–662.

[5] C. Lin, P. Hu, H. Su, S. Li, J. Mei, J. Zhou, and H. Leung, "Sensemood: depression detection on social media," in *Proceedings of the 2020 international conference on multimedia retrieval*, 2020, pp. 407–411.

[6] G. Shen, J. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T.-S. Chua, W. Zhu *et al.*, "Depression detection via harvesting social media: A multimodal dictionary learning solution." in *IJCAI*, 2017, pp. 3838–3844.

[7] M. A. Moreno, L. A. Jelenchick, K. G. Egan, E. Cox, H. Young, K. E. Gannon, and T. Becker, "Feeling bad on facebook: Depression disclosures by college students on a social networking site," *Depression and anxiety*, vol. 28, no. 6, pp. 447–455, 2011.

[8] J. H. Seah and K. J. Shim, "Data mining approach to the detection of suicide in social media: A case study of singapore," in *2018 IEEE international conference on big data (Big data)*. IEEE, 2018, pp. 5442–5444.

[9] A. U. Hassan, J. Hussain, M. Hussain, M. Sadiq, and S. Lee, "Sentiment analysis of social networking sites (sns) data using machine learning approach for the measurement of depression," in *2017 international conference on information and communication technology convergence (ICTC)*. IEEE, 2017, pp. 138–140.

[10] C. Zucco, B. Calabrese, and M. Cannataro, "Sentiment analysis and affective computing for depression monitoring," in *2017 IEEE international conference on bioinformatics and biomedicine (BIBM)*. IEEE, 2017, pp. 1988–1995.

[11] M. Birjali, A. Beni-Hssane, and M. Erritali, "Machine learning and semantic sentiment analysis based algorithms for suicide sentiment prediction in social networks," *Procedia Computer Science*, vol. 113, pp. 65–72, 2017.

[12] H. Almeida, A. Briand, and M.-J. Meurs, "Detecting early risk of depression from social media user-generated content." in *CLEF (working notes)*, 2017.

[13] T. Al Hanai, M. M. Ghassemi, and J. R. Glass, "Detecting depression with audio/text sequence modeling of interviews." in *Interspeech*, 2018, pp. 1716–1720.

[14] S. Madhu, "An approach to analyze suicidal tendency in blogs and tweets using sentiment analysis," *Int. J. Sci. Res. Comput. Sci. Eng*, vol. 6, no. 4, pp. 34–36, 2018.

[15] G. Coppersmith, R. Leary, P. Crutchley, and A. Fine, "Natural language processing of social media as screening for suicide risk," *Biomedical informatics insights*, vol. 10, p. 1178222618792860, 2018.

[16] J. J. Stephen and P. Prabu, "Detecting the magnitude of depression in twitter users using sentiment analysis," *International Journal of Electrical and Computer Engineering*, vol. 9, no. 4, p. 3247, 2019.

[17] 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 Systems with Applications*, vol. 133, pp. 182–197, 2019.

[18] D. B. Victor, J. Kawsher, M. S. Labib, and S. Latif, "Machine learning techniques for depression analysis on social media-case study on bengali community," in *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. IEEE, 2020, pp. 1118–1126.

[19] S. Sharma and S. Sharma, "Analyzing the depression and suicidal tendencies of people affected by covid-19's lockdown using sentiment analysis on social networking websites," *Journal of Statistics and Management Systems*, vol. 24, no. 1, pp. 115–133, 2021.

[20] J.-W. Baek and K. Chung, "Context deep neural network model for predicting depression risk using multiple regression," *IEEE Access*, vol. 8, pp. 18 171–18 181, 2020.

[21] F. Barbieri, J. Camacho-Collados, L. Neves, and L. Espinosa-Anke, "Tweeteval: Unified benchmark and comparative evaluation for tweet classification," *arXiv preprint arXiv:2010.12421*, 2020.

[22] R. Sawhney, H. Joshi, S. Gandhi, and R. Shah, "A time-aware transformer based model for suicide ideation detection on social media," in *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, 2020, pp. 7685–7697.

[23] 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.

[24] C. Li, C. Braud, and M. Amblard, "Multi-task learning for depression detection in dialogs," *arXiv preprint arXiv:2208.10250*, 2022.

[25] S. Ghosal and A. Jain, "Depression and suicide risk detection on social media using fasttext embedding and xgboost classifier," *Procedia Computer Science*, vol. 218, pp. 1631–1639, 2023.

[26] A.-M. Bucur, A. Cosma, P. Rosso, and L. P. Dinu, "It's just a matter of time: Detecting depression with time-enriched multimodal transformers," *arXiv preprint arXiv:2301.05453*, 2023.

[27] 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.

[28] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, and L. Neves, "TweetEval: Unified benchmark and comparative evaluation for tweet classification," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1644–1650. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.148