A BERT-Based Model for Detecting Depression in Diabetes-Related Social Media Posts

Rdouan Faizi, Bouchaib Bounabat and Mahmoud El Hamlaoui
ENSIAS, Mohammed V University in Rabat, Morocco

Keywords: Digital Health, Diabetes, Depression, Social Media Analysis, NLP, BERT.

Abstract: This paper introduces a BERT-based model for detecting depression in diabetic social media posts. Based on transformer-based language models, the proposed approach is specifically designed to capture the specific linguistic patterns that are indicative of depressive symptoms. The model was trained on a dataset of comments retrieved from diabetes-related YouTube channels, which were then manually annotated as either 'Depression' or 'Well-being'. Through extensive experimentation, the model achieved a high classification accuracy of 93% on the test set. These findings highlight its potential as an effective tool for automated mental health monitoring in at-risk populations, particularly those coping with chronic health conditions such as diabetes.

1 INTRODUCTION

Depression has recently become one of the most pressing public health issues given its growing prevalence and substantial socio-economic impact on both individuals and society (Marwaha et al., 2023; Hassan et al., 2021). Depression is a mood disorder that is characterized by persistent feelings of sadness, hopelessness, and loss of interest or pleasure in previously enjoyable activities, and often leads to changes in appetite, sleep patterns, and cognitive function (Schulz, 2020; Chand et al., 2021). According to the World Health Organization (WHO), depression affects over 280 million people worldwide. Consequently, it is considered as a major cause of disability and plays a critical role in the global disease burden (WHO 2023; Dawood Hristova & Pérez-Jover, 2023).

In recent years, the rising proliferation of social media platforms such as Twitter, Facebook, YouTube and Reddit has changed how people express their opinions, thoughts and emotions (Faizi et al., 2017). These online platforms have become optimal virtual communities where individuals freely share personal experiences, engage in discussions, and seek support from others. The massive amounts of user-generated textual data on these platforms, thus, offer an ideal opportunity to use Natural Language Processing (NLP) and machine learning techniques to identify patterns that indicate mental health issues, such as depression from these social media posts.

Therefore, the automatic detection of depression from social media posts offers great potential for early intervention and support (Coppersmith, 2017). In fact, by analysing the linguistic cues, underlying sentiment expressions and behavioural patterns embedded within textual content, machine learning models can potentially identify individuals at risk of depression before symptoms escalate and things get worse (Guntuku et al, 2017). This proactive approach does not only facilitate timely mental health interventions but can also enhance the effectiveness of existing support systems by customizing interventions that meet individual needs.

The main objective of the present work is to propose a machine learning model that can automatically detect depression in social media textual data. However, we will be restricted to detecting depression among individuals with diabetes given the significant correlation obtaining between both states. In this respect, several studies have confirmed that diabetes patients usually experience a higher risk of depression due to the chronic nature of their illness, the continual management required, and the potential complications that are likely to arise. Accordingly, addressing depression amongst diabetes patients can result in better overall health outcomes, a higher quality of life, and more efficient diabetes management.

In this study, the machine learning approach that we propose for the task of depression detection is based on BERT (Bidirectional Encoder

Faizi, R., Bounabat, B., El Hamlaoui and M.
A BERT-Based Model for Detecting Depression in Diabetes-Related Social Media Posts.
DOI: 10.5220/0013636100003967
In *Proceedings of the 14th International Conference on Data Science, Technology and Applications (DATA 2025)*, pages 627-632
ISBN: 978-989-758-758-0; ISSN: 2184-285X

Representations from Transformers). Our choice of this pre-trained language model is driven by its capability to understand the context of words within a sentence through bidirectional training. This makes it well-suited for analysing the intricacies of the language used in social media posts.

The remainder of this paper is structured as follows. Section 2 reviews previous research on the automatic detection of depression. Section 3 introduces our BERT-based model for identifying depression in social media posts, especially among individuals with diabetes. Finally, Section 4 highlights the key findings and conclusions.

2 RELATED WORK

The detection of depression from textual data has been an active research area in the last couple of years. In this respect, various studies have explored a wide range of approaches, from traditional machine learning algorithms to advanced deep learning models, to enhance detection accuracy.

Our purpose in this section is to highlight the different approaches that have been opted for in the existing studies and shed light on their key findings. By analysing these works, we aim to identify effective strategies and potential areas for improvement in depression detection.

One of the earliest studies in depression detection was undertaken by Nadeem (2016). In his study devoted to Major Depressive Disorder, the author used a crowdsourced dataset of Twitter users who publicly admitted being diagnosed with depression. The author utilized a Bag of Words approach to quantify each tweet and applied several statistical classifiers. Yet, the findings of the study demonstrated that the Naive Bayes (NB) approach scored the highest out of all our classifiers with a ROC AUC score of 0.94.

In another research work, Shen et al. (2017) opted for a multimodal depressive dictionary learning model to detect depression through social media data. In this vein, the authors constructed a labelled dataset of depressed and non-depressed users and extracted six feature groups encompassing clinical depression criteria and online behaviours. Their findings revealed that the proposed approach significantly outperforms several baseline models by 3% to 10%, which demonstrates its effectiveness in detecting depressive behaviour.

For their parts, Stankevich et al. (2018) explored different sets of features for the task of depression detection based on the CLEF/eRisk 2017 dataset. In fact, they assessed different feature engineering techniques such as TF-IDF, word embeddings and bigrams together with machine learning models, namely Support Vector Machine (SVM) and Random Forest (RF). After a series of experiments, the SVM model was found to achieve a maximum F1-score of 63%, while the embedding model showed a high recall of 84.61% with a decent F1-score of 61.53%.

In an additional research study, Febriansyah et al. (2023) used posts from the Dreaddit dataset, sourced from Reddit, and tested various traditional machine learning models and text representation techniques. Specifically, they employed SVM, NB, Decision Tree (DT), and RF. Moreover, they leveraged Bag of Words and TF-IDF as text representation methods. Among the approaches they tested, SVM emerged as the most effective as it achieved an F1-score of 80%, an accuracy of 75%, a recall of 92%, and a precision of 71%.

In their attempt to detect depression, Bokolo and Liu (2023) fed a Twitter dataset into different models, namely Logistic Regression (LR), Bernoulli NB, RF, DistilBERT, SqueezeBERT, DeBERTa, and RoBERTa. The latter model (i.e. RoBERTa) was associated with the highest performance with an accuracy of 98.1% and a mean accuracy of 0.97 across 10 cross-validation folds.

Following the same line of inquiry, Vasha et al. (2023) used six machine learning classifiers, namely NB, SVM, RF, DT, LR, and K-Nearest Neighbor to identify depressive posts from Bangla-language social media texts. After evaluating the classifiers based on accuracy, precision, recall and F1 score, SVM emerged as the best-performing model as it demonstrated the highest accuracy in distinguishing depressive content, followed by RF and LR. The SVM classifier achieved an accuracy of 75%, with a precision of 0.77, recall of 0.73, and an F1 score of 0.75.

To classify users into healthy, depressed, or at risk of self-harm, Naseem et al. (2023) introduced an emotion and time-aware architecture for detecting mental health conditions from social media posts. Unlike conventional methods that focus mainly on recent posts, their model considers users' historical emotional context and posting patterns. The results of the experiments that the authors conducted demonstrated that the proposed approach achieved F1-scores of 0.69 for depression and 0.62 for self-harm detection.

Unlike most previous studies which are primarily based on conventional machine learning models, Chen et al. (2023) proposed a hybrid deep learning model that combines pre-trained Sentence BERT (SBERT) for semantic representation learning with a Convolutional Neural Network (CNN) for temporal pattern identification. Using the Self-Reported Mental Health Diagnoses (SMHD) dataset, the

proposed SBERT-CNN achieved an accuracy of 0.86 and an F1 score of 0.86.

Most recently, Lamichhane (2023) evaluated the performance of ChatGPT using the GPT-3.5-turbo backend in three mental health classification tasks: stress detection (2-class), depression detection (2-class), and suicidality detection (5-class). By using annotated social media posts from publicly available datasets, the author employed ChatGPT's API for zero-shot classification based on specific prompts. Results indicated F1 scores of 0.73, 0.86, and 0.37 for stress, depression, and suicidality detection, respectively.

In conclusion, the reviewed studies demonstrate that various approaches have been employed to detect depression from textual data. These research works clearly illustrate the evolution of techniques, from traditional machine learning to advanced deep learning models, and the ongoing efforts to improve detection accuracy and effectiveness in detecting depressive symptoms from social media posts.

3 PROPOSED APPROACH

The approach we propose to detect depression in diabetes-related social media posts uses a BERT-based model. Our aim in this section is to detail the different stages that are involved in the methodology we adopted. This includes data preparation, model architecture, experimental setup, results and comparative analysis.

3.1 Data Preparation

The dataset used in this study consists of 11.860 user comments extracted from diabetes-related YouTube channels. To ensure its quality and relevance, the data underwent a rigorous cleaning process to remove irrelevant information, such as special characters, URLs, and non-informative text. Subsequently, the resulting comments were manually labeled into two classes: Depression, which includes comments expressing sadness, hopelessness, or distress (6.300 comments), and Well-being, which is composed of comments that denote positive emotions (5.100 comments). Neutral comments (i.e. 460) that do not express clear depressive or well-being feelings were excluded from our dataset. The table below provides a clear illustration of user comments classified into 'depression' and 'well-being' based on their emotional content.

Table 1: A sample of Diabetes Comments.

Comment	Class
Having diabetes is a struggle but I am happy t...	well-being
I have diabetes type 2 I know it hard to go th...	depression
I got diagnosed with Type 1 a year ago I'm 26....	depression
I hope diabetic patients can have healthy and ...	depression
No carbs no sugar and exercise. You'll be alri...	well-being

After data extraction, cleaning, and labeling, the next step involves data preprocessing. This includes splitting, tokenization and encoding. Initially, the data is split into two sets: 80% for training and 20% for testing. The training set is used to train the model and the test set is used to evaluate the model's performance on unseen data. Next, the data in each split is tokenized using the Bert tokenizer. This involves converting the text into tokens or individual words (or sub-words) that the model can understand. Then, the tokenized text is encoded into input IDs and attention masks. The former represents the tokenized words, while the latter enable the model to distinguish between actual and padding tokens.

3.2 Model Architecture

The BERT-based model proposed in this study is composed of various components that work together to process and classify comments. The architecture of this model is illustrated below.

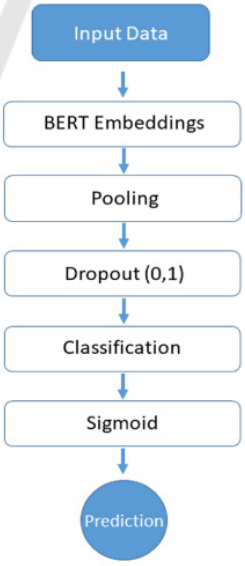

Figure 1: Architecture of the proposed model.

As the figure above shows, the input data, which has been tokenized and encoded, is fed into the BERT

model (uncased). The latter processes these inputs to generate contextual embeddings for each token. These embeddings, which capture the complex relationships and semantic meanings within the text, go through a pooling layer that aggregates them into a fixed-size vector that represents the entire sequence. This vector is passed through a dropout layer with a rate of 0.1 to prevent overfitting during training and the output is subsequently fed into a linear classification layer which applies a linear transformation to generate logits. The latter represent the scores that indicate the probability of each class (i.e. 'Depression' or 'Well-being'). Finally, a sigmoid function is applied to these logits to produce probabilities. The class associated with the highest probability is selected as the predicted label.

3.3 Experimental Setup

The training of the BERT-based model we proposed is carried out using the Adam optimizer, which has proven to be effective for fine-tuning transformer models. A learning rate of 1e-5 is selected to ensure that the model learns gradually and avoids overfitting to the training data. The binary cross-entropy loss function is used as it is well-suited for binary classification tasks, given that our objective is to distinguish between two classes, namely “depression” and “well-being”.

The model is trained for 3 epochs with a batch size of 64 samples per iteration. This provides sufficient exposure to varied examples while maintaining efficiency. To prevent overfitting, a dropout rate of 0.1 is applied during training. Once the training phase is over, the model is evaluated on a separate test set that was not used in the training process. This guarantees an unbiased assessment of the model's generalization ability.

To evaluate the model's performance, accuracy, precision, recall, and F1-score were made use of. These metrics provide a comprehensive view of the model's ability to classify comments into the correct categories. Accuracy measures the overall correctness of the model, while precision and recall provide insights into how well the proposed model is able to distinguish between the two categories. The F1-score, which is the harmonic mean of precision and recall, is particularly useful for gauging balance between the two classes. These evaluation results are used not only to assess the model's current performance, but also to identify potential areas for improvement, especially in achieving a more balanced classification of both “depression” and “well-being” comments.

3.4 Results and Discussion

Upon completing the evaluation process, the results are compiled and presented in the table below.

Table 2: Performance of the Bert-based model.

Metric	Depression	Well-being	Average
Precision	0.88	0.98	0.93
Recall	0.98	0.90	0.94
F1-Score	0.93	0.94	0.93
Accuracy	0.93		

By achieving an accuracy of 0.93 on the test set, the model demonstrates a high level of overall performance. This accuracy score clearly confirms the model's effectiveness in correctly classifying the majority of instances in the test data. These findings specifically indicate that the model excels in predicting instances of the 'Depression' class given its high recall, which is crucial for ensuring that most cases of depression are accurately identified. Nevertheless, the model is slightly less effective at predicting the 'Well-being' class, which denotes a potential area for improvement in balancing the model’s performance across both classes.

When compared to previously reviewed studies on depression detection, our BERT-based model delivered better results across several key evaluation metrics. Consult the following table:

Table 3: Comparative analysis of performance metrics across different works.

Study	**Models**	**Performance (Best Model)**
Nadeem (2016)	Naive Bayes (NB)	ROC AUC: 0.94 (NB)
Shen et al. (2017)	Multimodal	Outperforms baselines by 3%-10%
Stankevich et al. (2018)	SVM, Random Forest (RF)	SVM: F1-score 63%; Embeddings: Recall 84.61%, F1-score 61.53%
Febriansyah et al. (2023)	SVM, NB, Decision Tree (DT), RF	SVM: F1-score 80%, Accuracy 75%, Recall 92%, Precision 71%
Bokolo & Liu (2023)	LR, Bernoulli NB, RF, DistilBERT, SqueezeBERT, DeBERTa, RoBERTa	RoBERTa: Accuracy 98.1%, Mean Accuracy 97% (10-fold CV)
Vasha et al. (2023)	NB, SVM, RF, DT, LR, KNN	SVM: Accuracy 75%, Precision 0.77, Recall 0.73, F1-score 0.75

Table 3: Comparative analysis of performance metrics across different works (cont.).

Study	Models	Performance (Best Model)
Naseem et al. (2023)	Custom Deep Learning Model	Depression F1-score: 0.69, Self-harm F1-score: 0.62
Chen et al. (2023)	SBERT - CNN	Accuracy 0.86, F1-score 0.86
Lamichhane (2023)	GPT-3.5-turbo	F1-scores: Stress 0.73, Depression 0.86, Suicidality 0.37
Ours (2024)	BERT	Accuracy: 93%, Precision: 93%, Recall: 94%, F1-score: 94%

As the table above shows, the proposed BERT-based model confirms its superiority via the key metrics of accuracy, precision, recall and F1-score.

As far as precision is concerned, our model achieved a good score of 93%, thus, outperforming traditional models. For instance, the SVM method used by Febriansyah et al. (2023) recorded a precision of 0.71, while the SVM model employed by Vasha et al. (2023) performed slightly better at 0.77. The high precision achieved by our model reflects its ability to correctly identify relevant cases while minimizing false positives.

In terms of recall, our model's rate of 94% surpasses those of different approaches put forward in many other studies. For example, Febriansyah et al. (2023) achieved a praiseworthy recall of 0.92 with their SVM approach, while Stankevich et al. (2018) reported a recall of 84.61% using a word embedding model. Our BERT model's strong recall highlights its capability to effectively capture depressive signals. This ensures that very few relevant cases are missed. When associated with high precision, this robust recall contributes to an overall strong F1-score of 93%, which greatly reflects the model's balanced performance.

Regarding the F1-score, our model achieved an average of 93%. This result reflects a well-balanced performance in both precision and recall. This score surpasses the 0.86 reported by both Chen et al. (2023) using their SBERT-CNN model and Lamichhane (2023) as well as the 0.86 obtained by Febriansyah et al. (2023) with their SVM model. This proves our model's ability to maintain a strong balance between identifying relevant cases and minimizing false positives. The F1-score of the proposed BERT-based model proves its capacity to classify both 'depression' and 'well-being', which illustrates its potential across the mental health spectrum.

As for accuracy, our model reached an impressive score of 93%. Therefore, it is among the top performers in depression detection tasks. This accuracy surpasses the 0.86 achieved by Chen et al. (2023), who used a hybrid deep learning model, while traditional models from Febriansyah et al. (2023) and Vasha et al. (2023) achieved accuracies of around 75%. The high accuracy of 93% reflects our model's strong integration of precision, recall, and F1-score, which yields a consistent performance across various metrics. Although models like Bokolo & Liu (2023) with RoBERTa attained a higher accuracy of 98%, they operated within a different context and dataset. Nevertheless, our BERT model's performance across these interconnected metrics reinforces its robustness and reliability in identifying both 'depression' and 'well-being'.

Based on the analysis above, it is quite clear that the proposed model demonstrates its efficiency in depression detection given its high performance across multiple evaluation metrics. The interconnections among precision, recall, F1-score, and accuracy highlight the model's capability in effectively balancing the identification of true positive cases while minimizing false positives. Compared to traditional and even some contemporary deep learning models, our findings confirm the strength of transformer-based architectures in addressing mental health classification tasks. These findings not only contribute to the ongoing research in this field, but also hold promise for practical applications in mental health assessment and support.

4 CONCLUSION

The objective of this work was to propose a BERT-based model for detecting depression in diabetes-related social media posts. By harnessing the capabilities of transformer-based language models, the proposed approach effectively managed to capture the linguistic patterns indicative of depressive symptoms. In fact, it demonstrated a high level of accuracy at 93% in classifying comments as either 'Depression' or 'Well-being'. However, though the model showed minor limitations in predicting comments belonging to the latter class, it successfully managed in spotting 'Depression' instances with high precision and recall. This performance confirms the potential of the model as a reliable tool for detecting depressive symptoms in social media posts. Accordingly, it is likely to offer valuable support for mental health monitoring in online communities. In

future work, we will focus on enhancing the model's consistency in classifying well-being comments by using larger datasets and addressing class imbalances.

ACKNOWLEDGEMENTS

This work was supported by the AÏDA – Artificial Intelligence for DiAbetes project, under the AL-KHAWARIZMI programme. The authors gratefully acknowledge the financial and institutional support provided for the successful completion of this research.

REFERENCES

Bokolo, B. G., & Liu, Q. (2023). Deep learning-based depression detection from social media: Comparative evaluation of ml and transformer techniques. Electronics, 12(21), 4396.

Chand, S. P., Arif, H., & Kutlenios, R. M. (2021). Depression (nursing). Retrieved from https://europepmc.org/article/nbk/nbk568733

Chen, Z., Yang, R., Fu, S., Zong, N., Liu, H., & Huang, M. (2023, June). Detecting Reddit users with depression using a hybrid neural network SBERT-CNN. In 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI) (pp. 193-199). IEEE.

Coppersmith, G., Dredze, M., Harman, C., & Hollingshead, K. (2015). From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and clinical psychology: from linguistic signal to clinical reality* (pp. 1-10).

Dawood Hristova, J. J., & Pérez-Jover, V. (2023). Psychotherapy with psilocybin for depression: systematic review. Behavioral Sciences, 13(4), 297.

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Faizi R., El Fkihi S., El Afia A. & Chiheb R. (2017). Extracting Business Value from Big Data. In Proceedings of the 29th International Business Information Management Association Conference (IBIMA). ISBN: 978-0-9860419-7-6. 3-4 May 2017, Vienna, Austria

Febriansyah, M. R., Yunanda, R., & Suhartono, D. (2023). Stress detection system for social media users. Procedia Computer Science, 216, 672-681.

Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, *18*, 43-49.

Hassan, N. M., Kassim, E. S., & Said, Y. M. U. (2021). Financial wellbeing and mental health: a systematic review. Studies of Applied Economics, 39(4).

Lamichhane, B. (2023). Evaluation of ChatGPT for NLP-based mental health applications. arXiv preprint arXiv:2303.15727.

Marwaha, S., Palmer, E., Suppes, T., Cons, E., Young, A. H., & Upthegrove, R. (2023). Novel and emerging treatments for major depression. The Lancet, 401(10371), 141-153.

Nadeem, M. (2016). Identifying depression on Twitter. arXiv preprint arXiv:1607.07384.

Naseem, U., Thapa, S., Zhang, Q., Rashid, J., Hu, L., & Nasim, M. (2023, November). Temporal tides of emotional resonance: A novel approach to identify mental health on social media. In Proceedings of the 11th International Workshop on Natural Language Processing for Social Media (pp. 1-8).

Rehmani, F., Shaheen, Q., Anwar, M., Faheem, M., & Bhatti, S. S. (2024). Depression detection with machine learning of structural and non-structural dual languages. Healthcare Technology Letters.

Schulz, D. (2020). Depression development: From lifestyle changes to motivational deficits. Behavioural Brain Research, 395, 112845.

Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., ... & Zhu, W. (2017, August). Depression detection via harvesting social media: A multimodal dictionary learning solution. In IJCAI (pp. 3838-3844).

Stankevich, M., Isakov, V., Devyatkin, D., & Smirnov, I. V. (2018, January). Feature engineering for depression detection in social media. In ICPRAM (pp. 426-431).

Tejaswini, V., Sathya Babu, K., & Sahoo, B. (2024). Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model. ACM Transactions on Asian and Low-Resource Language Information Processing, 23(1), 1-20.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS)* (pp. 5998-6008). Retrieved from https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

WHO (2023). *Depression*. Retrieved April 10, 2025, from https://www.who.int/health-topics/depression