# Enhancing Depression Detection using BERT Models Pre-trained on Reddit Corpora

**Yuan Gao**

fmw4328@autuni.ac.nz

## Abstract

Depression is a pervasive and severe mental health disorder affecting millions worldwide, with its often covert nature making early detection challenging (World Health Organization, 2021). The proliferation of social media platforms, particularly Reddit, has created unprecedented opportunities for individuals to express their mental health concerns and seek support online (De Choudhury & De, 2014). This digital footprint provides a unique avenue for leveraging natural language processing techniques to automatically identify users potentially suffering from depression, facilitating early intervention. This study builds upon the model architecture proposed by Chen et al. (2023), which utilizes BERT (Bidirectional Encoder Representations from Transformers)(Devlin et al., 2019) for feature extraction from individual user posts, followed by a Convolutional Neural Network(Krizhevsky, Sutskever, & Hinton, 2017) for user-level classification. While this approach has shown promise, we hypothesize that the pre-trained BERT model, typically trained on formal corpora such as books and Wikipedia(Devlin et al., 2019), may not optimally capture the nuanced language patterns prevalent in social media discourse. To address this potential limitation, we propose a novel approach of pre-training the BERT model on a large corpus of Reddit data before integrating it into the BERT+CNN architecture. This study aims to evaluate whether this Reddit-specific pre-training can enhance the model's performance in detecting depression through social media content analysis. We conducted extensive experiments comparing the performance of the original BERT+CNN model against our Reddit-pre-trained variant. Performance metrics including accuracy, recall, F1 score, and validation loss were meticulously analyzed. Our findings indicate a significant improvement in performance, with the Reddit-pre-trained model achieving a 2.1 point increase in F1 score compared to the baseline model. This research contributes to the growing body of literature on digital mental health assessment and demonstrates the potential of domain-specific language model pre-training in improving the accuracy of depression detection in social media contexts. The implications of this study extend to both clinical practice and public health policy, offering insights into more effective, data-driven approaches for early mental health intervention strategies.

**Keywords:** *BERT; Convolutional Neural Network; CNN; Deep learning; Natural language processing; Text classification; Mental health; Depression; Social media; Reddit; SMHD*

# I. Introduction

Depression is a common and serious mental health disorder that significantly impacts individuals' thoughts, emotions, and daily functioning. Characterized by persistent feelings of sadness, loss of interest or pleasure, and a range of physical and cognitive symptoms, depression affects an estimated 3.8% of the global population, including 5.0% of adults and 5.7% of adults older than 60 years (World Health Organization, 2021). This prevalence translates to more than 280 million people worldwide living with depression, underscoring its status as a major public health concern. The impact of depression extends far beyond emotional distress. It is a leading cause of disability worldwide and contributes significantly to the global burden of disease (GBD 2019 Mental Disorders Collaborators, 2022). Depression can lead to various complications, including decreased productivity, strained relationships, and in severe cases, self-harm or suicide (American Psychiatric Association, 2013). The economic burden of depression is also substantial, with estimates suggesting that depression and anxiety disorders cost the global economy US$ 1 trillion each year in lost productivity (World Health Organization, 2021).

Despite its prevalence and severe consequences, depression often remains undiagnosed and untreated due to its covert nature. Many individuals experiencing depression may not recognize their symptoms or may be reluctant to seek help due to stigma associated with mental health issues (Corrigan et al., 2014). Additionally, the manifestation of depression can vary greatly among individuals, making it challenging for healthcare providers to identify and diagnose the condition in its early stages (Fried & Nesse, 2015). Early detection and intervention are crucial in managing depression effectively and preventing its progression to more severe states. Research has shown that early treatment can lead to better outcomes, reduced risk of recurrence, and improved quality of life (Ghio et al., 2014). However, the challenge lies in developing effective methods for early identification, particularly given the often subtle and variable presentation of depressive symptoms.

In recent years, there has been growing interest in leveraging technology and data analytics to aid in the early detection of depression. Particularly promising is the potential of analyzing digital footprints, such as social media activity, to identify patterns indicative of depressive symptoms (De Choudhury & De, 2014). This approach offers a unique opportunity to detect signs of depression in naturalistic settings, potentially before individuals themselves recognize the need for professional help. Researchers have employed various Natural Language Processing (NLP) techniques and text classification methods in an effort to improve performance. Early studies use single set features (such as bag of words(Nadeem, 2016, Paul, Jandhyala & Basu, 2018), N-grams(Benton, Mitchell & Hovy, 2017), LIWC(Coppersmith et al., 2015a) or LDA(Maupomé & Meurs, 2018, Resnik et al., 2015a)) or combinations of features (such as N-grams+LIWC(Wolohan et al., 2018) or BOW+LDA and TF-IDF+ LDA(Tyshchenko, 2018)) together with machine learning methods such as Logistic Regression, Support Vector Machine, Random Forest, Adaptive Boosting and Multilayer Perceptron classifier.

Later success in the use of deep learning for NLP tasks have motivated the identification of depression from social media posts by these deep learning techniques including Convolutional Neural Network (CNN) (Yates, Cohan & Goharian, 2017, Souza, Nobre & Becker, 2021), Recurrent Neural Networks (RNN)(Souza, Nobre & Becker, 2021, Souza, Nobre & Becker, 2020), and Transformers (Dinu & Moldovan, 2021, Jiang et al., 2020). Recent advancements in natural language processing, particularly the development of Transformer-based models like BERT, have significantly improved the capability to detect depression through social media data analysis. Tadesse et al. (2019) utilized BERT for depression detection on Twitter data, demonstrating improved performance over traditional machine learning methods. Their model achieved an F1-score of 0.93, showcasing BERT's effectiveness in capturing nuanced language patterns associated with depression. Martínez-Castaño et al. (2021) employed a BERT-based approach for early depression detection on Reddit. They fine-tuned BERT on the eRisk 2018 dataset, achieving competitive results with an ERDE5 score of 0.063, highlighting the model's ability to identify early signs of depression. Chen et al. (2023) developed a hybrid neural network that integrates sentence BERT and CNN for identifying Reddit users with depression. The sentence BERT allows the learning of meaningful representation of each post; CNN enables the further transformation of those embeddings via convolution operation for identification of depression patterns of users.

Despite the promising results achieved by BERT-based models in depression detection, there remain significant limitations in current approaches. A key challenge is the potential mismatch between the language used in pre-training BERT and the language commonly found on social media platforms. BERT's pre-training typically involves formal text sources such as Wikipedia and books (Devlin et al., 2019), which differ substantially from the informal, colloquial, and often unstructured language used on social media (Nguyen et al., 2020). This linguistic discrepancy can lead to suboptimal performance when applying pre-trained BERT models directly to social media data for depression detection. The unique vocabulary, syntax, and contextual nuances of social media communication may not be adequately captured by models trained on more formal text corpora. Consequently, there is a growing recognition of the need for domain-specific language model pre-training to better adapt these powerful models to the specific challenges of mental health detection on social media platforms (Gururangan et al., 2020).

Given these limitations, the main research question of this study is: Can pre-training BERT on Reddit data improve the accuracy of depression detection in social media contexts? To address this question, we propose an approach that involves pre-training the BERT model on a large corpus of Reddit data before integrating it into a depression detection framework.

Our approach consists of the following steps:

1. Collecting a diverse and representative dataset of Reddit posts, including both depression-related and general content.
2. Pre-training the BERT model on this Reddit-specific dataset.
3. Integrating the pre-trained BERT model into a depression detection pipeline, similar to the architecture proposed by Chen et al. (2023).

4. Evaluating the performance of this Reddit-pre-trained BERT model against baseline models (original BERT).

We hypothesize that this domain-specific pre-training will enable the model to better capture the nuanced language patterns associated with depression in social media contexts, potentially leading to improved detection accuracy and earlier identification of at-risk individuals.

The remainder of this work is divided into five sections. Section II introduces the relevant machine learning methods and social media datasets including depression labels. Section III describes the structure of our model and training process in details. Section IV shows the model performance and experiment results. Section V talks about future works. Section VI concludes this work.

## II. Related work

### A. Depression dataset

Social media platforms have emerged as a rich source of data for researchers investigating the linguistic characteristics of individuals with mental health conditions. Twitter, with its widespread popularity and short-form public messages, became an early focus of such studies. Researchers used crowdsourcing to collect gold standard labels on a cohort's depression and proposed a variety of social media measures to identify users potentially suffering from depression (De Choudhury et al., 2013). Interestingly, these studies revealed that certain language features were consistent across different cultures, as demonstrated by similar findings in both English and Japanese tweets (Tsugawa et al., 2015). This cross-cultural consistency suggests that some linguistic indicators of mental health conditions may be universal.

As the field progressed, researchers began to shift away from survey-based methods due to their cost and potential bias. Instead, they focused on analyzing the content shared by social media users to identify mental health conditions. Coppersmith et al. (2014) identified approximately 1,200 Twitter users with four mental health conditions (bipolar, depression, PTSD, SAD) using diagnosis statements found in tweets.

Although the numerous short texts on Twitter offer some understanding of the language characteristics of individuals with mental health conditions, long-form content can reveal further linguistic insights. Reddit, with its long-form content and diverse communities, provided a new avenue for investigation. Losada & Crestani (2016) applied the self-reported diagnosis strategy to identify approximately 150 Reddit users who suffer from depression. Later, Yates et al. (2017) developed the Reddit Self-reported Depression Diagnosis (RSDD) dataset, which included over 9,000 users with depression and 100,000 control users. Researchers also explored other data sources, including student essays (Resnik et al., 2013)

and text message conversations from mental health crisis centers (Althoff et al., 2016). These diverse sources allowed for a more comprehensive understanding of the language used by individuals with mental health conditions in various contexts. The shared task at the 2nd Computational Linguistics and Clinical Psychology Workshop (CLPsych 2015) focused on identifying depression and PTSD users on Twitter (Coppersmith et al., 2015b), with leading submissions relying on the LIWC lexicon, topic modeling, and other domain-dependent features (Resnik et al., 2015b; Preoţiuc-Pietro et al., 2015).

Cohan et al. (2018) build on and address key limitations of previous research to form SMHD. Similar to the Reddit Self-reported Depression Diagnosis (RSDD) dataset (Yates et al., 2017), they construct their corpus using self-reported diagnoses from Reddit, which gathers a substantial amount of long-form content unconstrained by character limits. This provides a more natural form of language, compared to the abbreviated text often seen on platforms like Twitter. In this work, we use SMHD as our train and validation dataset.

## B. Depression detection method

In 2018, Cohan et al. used traditional machine learning machine learning methods (such as Logistic Regression, SVM, and XGBoost), CNN, and FastText (Joulin 2016) to develop a binary classifier to identify individuals with depression. The best performance was achieved by supervised FastText with a F1 score of 0.54. (Cohan et al., 2018).

In 2019, Strube et al. employed the Hierarchical Attention Network for depression identification with SMHD dataset and achieved a F1 score of 0.68 (Sekulić & Strube, 2020).

In 2021, Dinu and Moldovan used three pretrained transformer models - BERT, XLNET, and RoBERTa to develop a post-level binary classifier using SMHD dataset and obtained a F1 score of 0.68, 0.70, and 0.68, respectively. (Dinu & Moldovan, 2021)

In 2022, Souza et al. applied word embedding techniques, such as GloVe and Word2Vec, to generate domain-specific embeddings for Reddit posts within the SMHD dataset. These embeddings were then utilized in CNN and LSTM models for depression detection, achieving F1 scores of 0.79 and 0.77, respectively. (Souza, Nobre & Becker, 2022)

In 2023, Chen et al. employed the SBERT-CNN model, which first utilizes SBERT (Reimers, 2019) to perform sentence-level feature extraction from Reddit users' posts. The sentence-level features for each user are then aggregated to obtain two-dimensional user-level features. A CNN model is subsequently trained to classify each user as either suffering from depression or as a control subject. The SBERT-CNN model achieved an accuracy of 0.86 and an F1 score of 0.86. (Chen et al., 2023)

## III. Methods

### A. Data construction

The Social Media Mental Health Dataset (SMHD) is a large-scale dataset designed for research in mental health detection using social media data. It contains user-generated posts from Reddit, specifically focusing on individuals who have self-reported mental health conditions. The dataset covers various mental health disorders, including depression, anxiety, bipolar disorder, schizophrenia, eating disorders, post-traumatic stress disorder (PTSD), and obsessive-compulsive disorder (OCD). SMHD includes both posts from diagnosed individuals and a control group, consisting of users who do not participate in mental health-related communities. This dataset is widely used for machine learning tasks, such as text classification and mental health detection, enabling research into early detection and intervention for mental health conditions based on social media activity.

In this study, we only use the depression data and control group data from the SMHD dataset while data related to other mental illnesses are discarded. The initial SMHD dataset is divided into three segments—training, validation, and testing. We preserved the original separation. The depression data used was the original one and the data for the control group was randomly sampled with the number of users being the same as depression group. Detailed information regarding the depression dataset we generated is presented in Table I.

Table I. Depression dataset from SMHD

| Datasets | Labels | Total Users | Total Posts |
|---|---|---|---|
| **Train** | Depression | 1316 | 216,022 |
| | Control | 1316 | 391,286 |
| **Valid** | Depression | 1308 | 290,858 |
| | Control | 1308 | 394,127 |
| **Test** | Depression | 1316 | 209,188 |
| | Control | 1316 | 393,545 |

### B. Data preprocessing

Given the diverse linguistic styles present on Reddit, a popular social media platform among younger demographics, rigorous preprocessing was essential to prepare the data. Our preprocessing pipeline consisted of several key steps designed to standardize the input and mitigate potential noise in the data.

First, we implemented a comprehensive cleaning process. All hyperlinks were removed from the text to eliminate extraneous information that could potentially skew our analysis. This step is crucial as links often do not contribute directly to the semantic content relevant for depression detection. Next, we addressed the prevalent use of emojis in social media

communication. Rather than simply removing these pictographs, we employed a semantic conversion technique. Each emoji was transformed into its corresponding textual description, preserving the emotional and contextual information they convey. This approach aligns with previous research suggesting that emoji usage patterns can be indicative of mental health states (Settanni & Marengo, 2015). To further standardize the input, we removed excessive line breaks and whitespace. This compression of redundant formatting elements served to reduce the overall token count without sacrificing meaningful content.

For posts exceeding the 512-token limit, a truncation procedure was applied. While this approach potentially results in the loss of some information, our analysis indicated that the most salient content for depression detection typically appears in the earlier portions of a post. Conversely, for posts falling short of the maximum length, we implemented zero-padding. This padding technique ensures uniformity across all samples within a batch, which is critical for efficient processing in many deep learning architectures.

### C. BERT+CNN pipeline

Inspired by Chen et al. (2023), who utilized fixed BERT for feature extraction from Reddit users' posts, then merged the one-dimensional embeddings of each post into two-dimensional features representing the user's features. Afterward, CNN was employed to train the extracted features to predict whether the user belongs to the depression or control group.

We adopted their pipeline structure (BERT+CNN). Assuming the number of users is $k$, and the number of posts published by the k-th user is $n_k$, all posts were padded or truncated to a fixed length of $512$ during preprocessing. We used the BERT base model for tokenization and sentence feature extraction. After processing through the BERT model, we obtained user-level features for each user, with dimensions of $768 \times n_k$, where $768$ is the length of the sentence-level embedding output by the BERT model for a single post. The input to the CNN is a two-dimensional matrix with a shape of $(N, 768)$, where $N$ represents the maximum number of posts per user and is set to $512$ in the experiment.

The CNN model is composed of two 1D convolutional blocks. Each block contains a convolutional layer, a max-pooling layer, and a dropout layer. The first convolutional layer uses a kernel size of 40x768, while the second convolutional layer has a kernel size of 40x32. The first and second convolutional layers utilize 32 and 16 filters, respectively. The ReLU activation function is applied in each convolutional layer to introduce non-linearity. Multiple convolutional layers are employed to extract higher-level features from the input matrix. Max-pooling layers are used to compute the maximum values from the feature map covered by the filters, reducing the dimensionality of the feature map by half. Both convolutional blocks have dropout rates of 20%. The flattening layer is fully connected to two hidden layers with 16 and 8 neurons respectively. The output layer consists of 2 neurons for binary classification, utilizing the Softmax loss function.

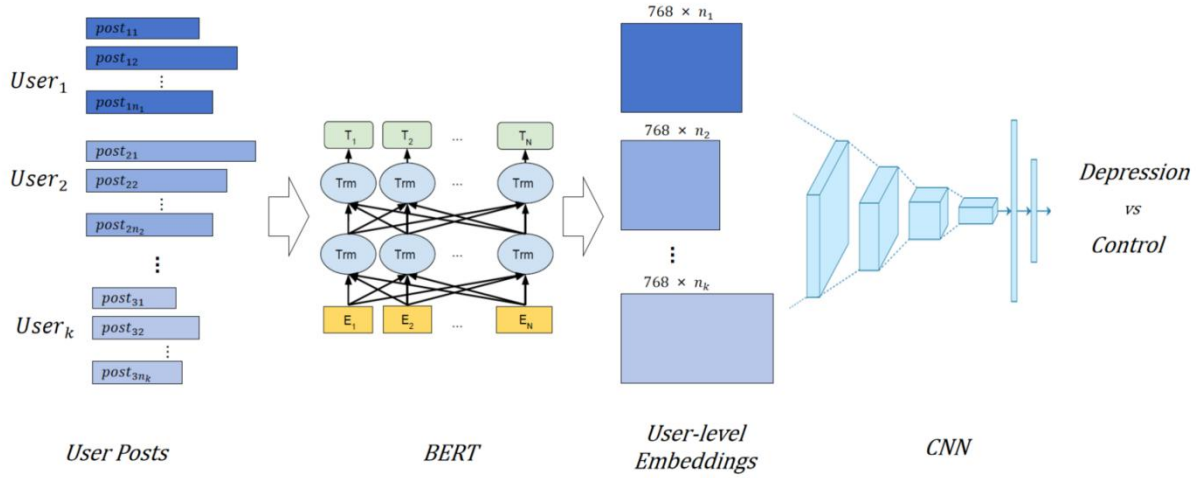


Figure I BERT+CNN pipeline, where k is the number of users, $n_k$ represents the the number of posts published by the k-th user

## D. BERT pre-training

BERT model's pre-training utilizes two methods:

1. **Masked Language Model (MLM)**: During BERT's pre-training, the model randomly selects 15% of the words in the input text and replaces them with a special mask token ([MASK]). The model's task is to predict the masked words. Since BERT is a bidirectional model, it can leverage both preceding and succeeding contextual information to predict the masked words, unlike traditional language models, which typically use unidirectional information. Through MLM, BERT is able to learn deeper semantic understanding.
2. **Next Sentence Prediction (NSP)**: NSP helps BERT learn the relationships between sentences. During training, the model is provided with two sentences. In 50% of the cases, the second sentence is the actual sentence that follows the first, while in the other 50%, it is a randomly chosen sentence. The model needs to predict whether the second sentence is the continuation of the first one. This task helps the model better understand contextual relationships in natural language.

As subsequent developments in BERT training have demonstrated that the MLM task is more important than NSP (Liu et al., 2019), this work, considering the constraints on computational resources, only pre-trained the model on the MLM task.

Training data and validation data are randomly selected from SMHD. In this work, 619,703 cleaned Reddit users' posts were trained, totaling around 24 million tokens. The batch size for each step was set to 16, and the model was trained for 1 epoch, resulting in a total of 38,732 steps. The validation loss over steps is shown in Figure II.

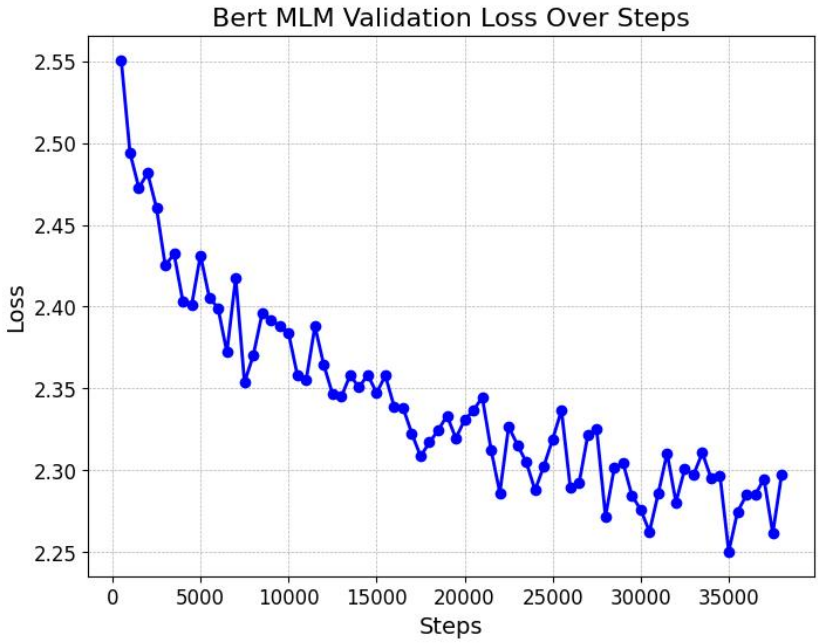


Figure II Bert pre-training loss over steps. The batch size for each step was 16.

The training utilized an NVIDIA T4 GPU on Google Colab, taking approximately 36 hours. The data preprocessing step was implemented before training. The implementation of BERT was carried out using the PyTorch framework (Paszke et al., 2019) and the Transformers library (Wolf et al., 2020).

### E. BERT+CNN fine-training

The pre-trained BERT and randomly initialized CNN were used for training, with all parameters being fine-tuned. The Adam optimizer is applied with a learning rate of 1e-4, and Crossentropy loss is used for parameter optimization. The model is trained and validated over 30 epochs. The CNN model requires a consistent size of input matrix (N, 768) for each user but the number of posts per user varies among users. In this work, the fixed post number N was set to 512. Any embeddings exceeding this length were discarded, while numbers smaller than 512 were padded with zeros.

## IV. Experiment results

Training with the original BERT model, the highest F1 score achieved was 0.813. After pretraining the BERT model, the highest F1 score increased to 0.834, showing an improvement of 0.021. The BERT+CNN pipeline with pretrained BERT model consistently had a lower training loss compared to the pipeline with original BERT model as shown below. Figure III shows the F1 score, accuracy, precision and train loss comparison between the

pipeline with original BERT (not pretrained on Reddit data) and pretrained BERT (this work) over training epochs. In the figure, smoothed lines (dotted lines) are drawn for better illustration.

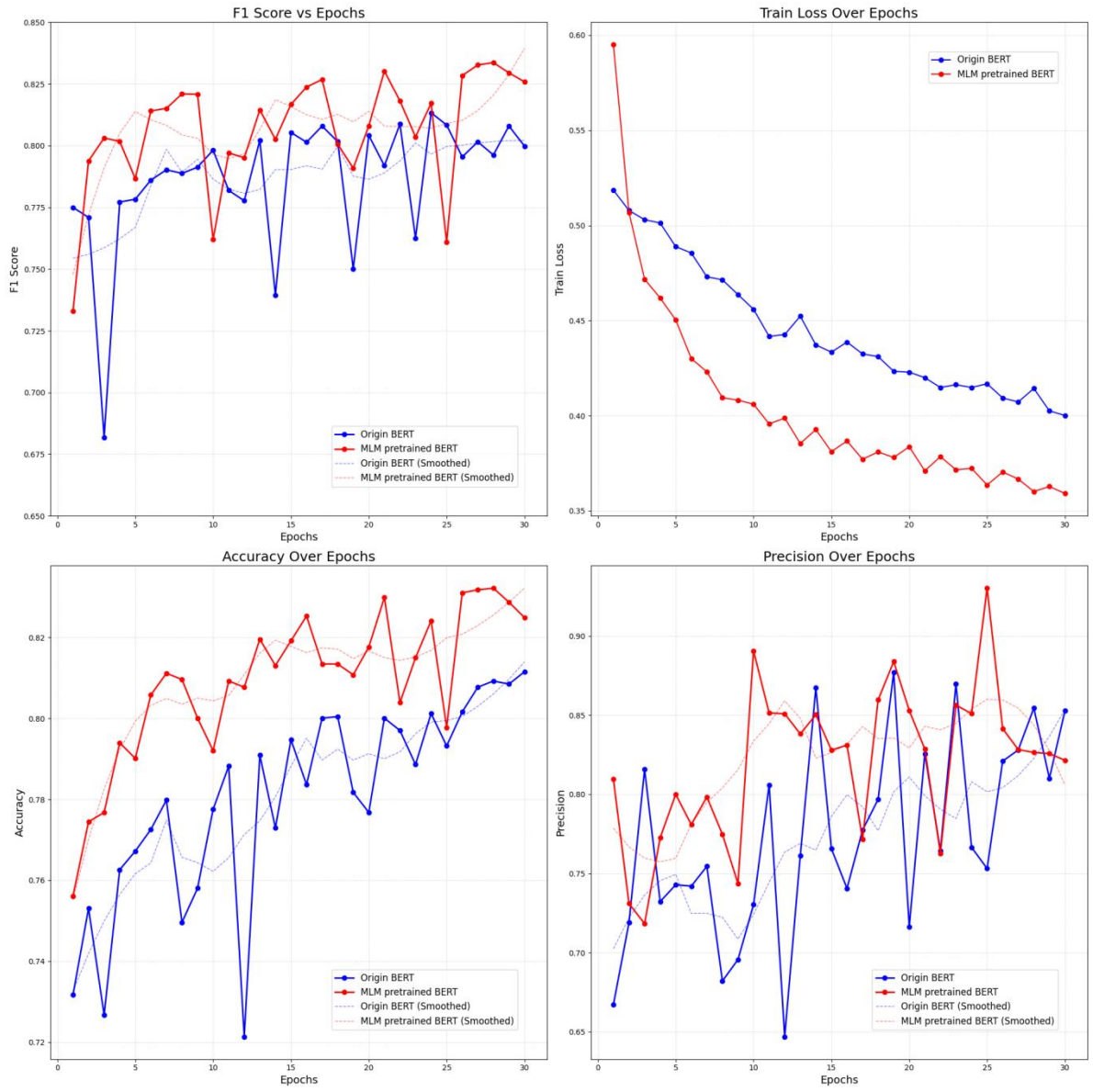


Figure III Comparison between BERT+CNN pipeline with original BERT and MLM pretrained BERT

Interestingly, in the first epoch, the fine-tuned BERT model performed worse than the original, with an F1 score lower by 0.034 and a training loss higher by 0.08. This is likely due to the fact that only the MLM task was used to pretrain BERT, and the NSP task was not used. As a result, the [CLS] token's embedding used for the downstream CNN task was randomly initialized, leading to inferior performance compared to the original BERT model in the early stages. None the less, from the second epoch onward, the pretrained BERT model consistently outperformed the original BERT model.

The experimental results demonstrate that using the pretrained BERT model for post feature extraction can significantly improve overall performance. Pre-training BERT models on domain-specific corpora can lead to significant improvements in performance on downstream tasks. This is because domain-specific pre-training allows the model to capture more nuanced and relevant linguistic patterns, terminologies, and contextual information that are characteristic of the target domain.

## V. Future works

While this study demonstrates the potential of domain-specific pre-trained BERT models in improving depression detection on the SMHD dataset, several areas remain for future exploration. One promising direction is the investigation of large language models(LLM), such as GPT-4(Achiam et al., 2023) or LLaMA(Touvron et al., 2023), which have demonstrated state-of-the-art performance across various NLP tasks. These models, due to their increased capacity and richer contextual understanding, may further enhance the accuracy and reliability of depression detection from social media data. Future work could focus on fine-tuning these larger models on mental health-specific corpora and comparing their performance with BERT-based models. Additionally, examining the interpretability of these large models in clinical contexts and their ethical implications for mental health diagnostics remains a critical area for future research.

## VI. Conclusion

This study demonstrates the significant impact of domain-specific pre-training on the performance of depression detection models in social media contexts. By pre-training the BERT model on a large corpus of Reddit data before integration into the BERT+CNN architecture, we achieved a notable 2.1 percentage point increase in F1 score compared to the baseline model. This improvement underscores the importance of capturing domain-specific linguistic nuances in mental health assessment tasks. Our findings reveal that while the pre-trained model initially underperformed, it consistently outpaced the original BERT model from the second epoch onwards. This observation highlights the potential long-term benefits of domain-specific pre-training, despite initial performance setbacks. The consistently lower training loss observed with the pre-trained model further supports its enhanced efficiency in feature extraction and classification. These results have important implications for both clinical practice and public health policy. They suggest that tailoring natural language processing models to the specific linguistic patterns of social media platforms can significantly enhance their efficacy in identifying potential mental health concerns. This approach offers a promising avenue for developing more accurate, data-driven early intervention strategies in mental health care. Future research should explore the integration of

additional pre-training tasks beyond Masked Language Modeling to further improve model performance. Moreover, investigating the generalizability of this approach to other mental health conditions and social media platforms could yield valuable insights for comprehensive digital mental health assessment strategies.

## References


Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774. https://doi.org/10.48550/arXiv.2303.08774

Althoff, T., Clark, K., & Leskovec, J. (2016). Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics, 4*, 463-476. https://doi.org/10.1162/tacl_a_00111

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Benton, A., Mitchell, M., & Hovy, D. (2017). Multi-task learning for mental health using social media text. arXiv preprint arXiv:1712.03538.

Chen, Z., Yang, R., Fu, S., Zong N., Liu, H., & Huang, M. (2023). Detecting Reddit users with depression using a hybrid neural network SBERT-CNN. *In Proceedings of the 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, 193-199. https://doi.org/10.1109/ICHI57859.2023.00035

Cohan, A., Desmet, B., Yates, A., Soldaini, L., MacAvaney, S., & Goharian, N. (2018). SMHD: a Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions. *In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), Proceedings of the 27th International Conference on Computational Linguistics,* 1485-1497. Association for Computational Linguistics.

Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying mental health signals in Twitter. *In Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality,* 51-60. https://doi.org/10.3115/v1/W14-3207

Coppersmith, G., Dredze, M., Harman, C., & Hollingshead, K. (2015a). 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,* 1-10. https://doi.org/10.3115/v1/W15-1201

Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. (2015b). CLPsych 2015 shared task: Depression and PTSD on Twitter. *In Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality,* 31-39. https://doi.org/10.3115/v1/W15-1204

Corrigan, P. W., Druss, B. G., & Perlick, D. A. (2014). The impact of mental illness stigma on seeking and participating in mental health care. *Psychological Science in the Public Interest, 15*(2), 37-70. https://doi.org/10.1177/1529100614531398

De Choudhury, M., & De, S. (2014). Mental Health Discourse on Reddit: Self-Disclosure, Social Support, and Anonymity. *Proceedings of the International AAAI Conference on Web and Social Media, 8*(1), 71-80. https://doi.org/10.1609/icwsm.v8i1.14526

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. *In Proceedings of the International AAAI Conference on Web and Social*

*Media, 7*(1), 128-137. https://doi.org/10.1609/icwsm.v7i1.14432

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),* 4171-4186. https://doi.org/10.18653/v1/N19-1423

Dinu, A., & Moldovan, A. C. (2021). Automatic detection and classification of mental illnesses from general social media texts. *In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, 358-366.

Fried, E. I., & Nesse, R. M. (2015). Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STAR*D study. *Journal of Affective Disorders, 172*, 96-102. https://doi.org/10.1016/j.jad.2014.10.010

GBD 2019 Mental Disorders Collaborators. (2022). Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet Psychiatry, 9*(2), 137-150. https://doi.org/10.1016/S2215-0366(21)00395-3

Ghio, L., Gotelli, S., Marcenaro, M., Amore, M., & Natta, W. (2014). Duration of untreated illness and outcomes in unipolar depression: A systematic review and meta-analysis. Journal of Affective Disorders, 175, 152-154, 45-51. https://doi.org/10.1016/j.jad.2013.10.002

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8342-8360. https://doi.org/10.18653/v1/2020.acl-main.740

Jiang, Z. P., Levitan, S. I., Zomick, J., & Hirschberg, J. (2020). Detection of mental health from reddit via deep contextualized representations. *In Proceedings of the 11th international workshop on health text mining and information analysis*, 147-156. https://doi.org/10.18653/v1/2020.louhi-1.16

Joulin, A. (2016). Fasttext. zip: Compressing text classification models. arXiv preprint arXiv:1612.03651.

Krizhevsky, A., & Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM, 60*(6), 84-90. https://doi.org/10.1145/3065386

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

Losada, D. E., & Crestani, F. (2016). A test collection for research on depression and language use. *In International conference of the cross-language evaluation forum for European languages*, 28-39.

Martínez-Castaño, R., Htait, A., Azzopardi, L., Moshfeghi, Y. (2021). BERT-based transformers for early detection of mental health illnesses. *In: Candan, K.S., et al. Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2021. Lecture Notes in Computer Science, 12880*, 189-200.

https://doi.org/10.1007/978-3-030-85251-1_15

Maupomé, D., & Meurs, M. J. (2018). Using Topic Extraction on Social Media Content for the Early Detection of Depression. *CLEF (working notes)*, 2125. https://api.semanticscholar.org/CorpusID:51942824

Nadeem, M. (2016). Identifying depression on Twitter. Retrieved from https://arxiv.org/abs/1607.07384

Nguyen, D. Q., Vu, T., & Nguyen, A. T. (2020). BERTweet: A pre-trained language model for English Tweets. *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 9-14. https://doi.org/10.18653/v1/2020.emnlp-demos.2

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Paul, S., Jandhyala, S. K., & Basu, T. (2018). Early Detection of Signs of Anorexia and Depression Over Social Media using Effective Machine Learning Frameworks. In CLEF (Working notes).

Preoţiuc-Pietro, D., Eichstaedt, J., Park, G., Sap, M., Smith, L., Tobolsky, V., ... & Ungar, L. (2015). The role of personality, age, and gender in tweeting about mental illness. *In Proceedings of the 2nd workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, 21-30. https://doi.org/10.3115/v1/W15-1203

Reimers, N. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv preprint arXiv:1908.10084. https://doi.org/10.18653/v1%2FD19-1410

Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V. A., & Boyd-Graber, J. (2015a). Beyond LDA: exploring supervised topic modeling for depression-related language in Twitter. *In Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, 99-107. https://doi.org/10.3115/v1/W15-1212

Resnik, P., Armstrong, W., Claudino, L., & Nguyen, T. (2015b). The University of Maryland CLPsych 2015 shared task system. *In Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, 54-60. https://doi.org/10.3115/v1/W15-1207

Resnik, P., Garron, A., & Resnik, R. (2013). Using topic modeling to improve prediction of neuroticism and depression in college students. *In Proceedings of the 2013 conference on empirical methods in natural language processing*, 1348-1353.

Settanni, M., & Marengo, D. (2015). Sharing feelings online: studying emotional well-being via automated text analysis of Facebook posts. *Frontiers in psychology, 6*, 1045.

Sekulić, I., & Strube, M. (2020). Adapting deep learning methods for mental health prediction on social media. arXiv preprint arXiv:2003.07634. https://doi.org/10.18653/v1/D19-5542

Souza, V., Nobre, J., & Becker, K. (2021). A deep learning ensemble to classify anxiety, depression, and their comorbidity from texts of social networks. *Journal of Information and Data Management, 12*(3). https://doi.org/10.1109/jbhi.2022.3151589

Souza, V., Nobre, J., & Becker, K. (2021). A Deep Learning Ensemble to Classify Anxiety. *Depression, and their Comorbidity from Texts of Social Networks*.

Souza, V. B., Nobre, J., & Becker, K. (2020). Characterization of anxiety, depression, and their comorbidity from texts of social networks. *Anais do XXXV Simpósio Brasileiro de Bancos de Dados*, 121-132.

Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of depression-related posts in Reddit social media forum. *IEEE Access, 7*, 44883-44893. https://doi.org/10.1109/ACCESS.2019.2909180

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971. https://doi.org/10.48550/arXiv.2302.13971

Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., & Ohsaki, H. (2015). Recognizing depression from twitter activity. *In Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 3187-3196. https://doi.org/10.1145/2702123.2702280

Tyshchenko, Y. (2018). Depression and anxiety detection from blog posts data. *Nature Precis. Sci., Inst. Comput. Sci., Univ. Tartu, Tartu, Estonia*.

Wolohan, J. T., Hiraga, M., Mukherjee, A., Sayyed, Z. A., & Millard, M. (2018). Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with NLP. *In Proceedings of the first international workshop on language cognition and computational models*, 11-21.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. *In Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 38-45.

World Health Organization. (2021). Depressive disorder (depression) https://www.who.int/news-room/fact-sheets/detail/depression

Yates, A., Cohan, A., & Goharian, N. (2017). Depression and self-harm risk assessment in online forums. *In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2968-2978.https://doi.org/10.18653/v1/D17-1322