Comparing the performance of pre-trained transformer models for depression symptom detection

Team 45

Abstract

This study compares the performance of BERT, RoBERTa, BERTTweet, and PHS-BERT in identifying symptoms in text through multi-label classification. Evaluation metrics include accuracy, precision, recall, F1-score, Exact Match Ratio, Hamming Loss, and Jaccard Score. BERT-base outperformed other models in most metrics, with BERTTweet excelling in Hamming Loss and Jaccard Score. Attention analysis offered insight into model behavior. Despite promising results, limitations like limited dataset and model selection require further research. Future work should focus on dataset expansion, alternative models, fine-tuning strategies, model interpretability, and real-world clinical applications.

1 Introduction

In recent years, the prevalence of mental health issues has become a growing concern worldwide. In the UK, around 1 in 6 persons suffer from depression, making it one of the most common mental health problems, and the World Health Organization (WHO) estimates depression to be the second most prevalent mental disorder worldwide (World Health Organization, 2022a). Since the start of the Covid-19 pandemic, depression has increased by 7%, and as of 2021, 17% of adults in the UK reported experiencing depression symptoms. (Office for National Statistics, 2021; World Health Organization, 2022b; Kumar and Nayar, 2021).

People with a depressive disorder are significantly impaired in their ability to function socially or occupationally in their daily life, and they are also at a very high risk of suicide (American Psychiatric Association, 2022). Therefore, it is extremely important to identify early signs of depression to be able to prevent the onset of it. An early assessment and treatment of potentially depressed people can curve the development of the underlying condition as well as alleviate the negative impacts of depression in their well-being and daily activities. Then, by identifying the common symptoms, healthcare professionals can use their domain knowledge to assess which combination of symptoms are potential early signs of depression, so they can intervene with the purpose of preventing the development of this mental condition.

Due to the anonymity offered by the internet, people have grown increasingly at ease talking about sensitive subjects like depression online. Additionally, social media is widely used to express and communicate feelings, emotions and thoughts, as well as to share personal experiences and daily struggles; therefore, social media is a rich source of information for early detection of depression symptoms which will aid medical specialists in recognizing early stages of depression or even combine with other sources of data to identify already depressed people that have been unnoticed by the health system.

The analysis of social media information using artificial intelligence (AI) and natural language processing (NLP) techniques can create new possibilities for early diagnosis, prevention, and treatment of mental health issues. Pre-trained BERT models, a state-of-the-art NLP model, have demonstrated an indisputable effectiveness and performance improvement in a wide variety of NLP tasks (Casola et al., 2022); however, domain-specific BERT models have shown even better performance than general pre-trained BERT models on target tasks that use in-domain data for pretraining (Dai et al., 2020). Therefore, in this paper, we aim to investigate the potential of using NLP techniques and specifically domain-specific transformer models compared to general-use transformer models to identify and analyse depressive symptoms through texts shared between users on social media. With our research, we hope to offer some new informa-

tion about this topic which can get us closer to understanding if and how, such techniques should or could be used in this context.

2 Related work

NLP has been used extensively in mental health detection tasks in the last decade; a recent systematic literature review on mental illness detection found an upward trend in NLP methods applied to mental illness detection research during the last couple of years, but in particular, deep-learning-based methods have increased in popularity since 2019 given their better performance compared to traditional machine learning methods (Zhang et al., 2022).

Depression is the mental illnesses that has attracted most interest in NLP mental health research, and research on social media constitutes around 81% of these publications (Zhang et al., 2022). However, most of the research related with depression has been done on diagnosing depression, but very few studies have tried to diagnose depression symptoms. Duvvuri et al. (2022) were one of those articles in which the aim of the study was to predict seven key symptoms for depression from chats on the social media platform Discord. Their approach was to create a prediction model for each symptom using the following models: Random Forest, Support Vector Machine, Naïve Bayes, and Convolution Neural Network. Finally, given that the Random Forest models outperformed the other models for each symptom, they built an ensemble Random Forest model based on the 7 symptom classifiers to predict depression. With these models, they obtained at least 73% accuracy in diagnosing depression symptoms and 99% accuracy in identifying depression. To test our hypothesis, we used the data collected during the above study; see more on 3.3.1.

Our approach to this task was different since we have not detected each depression symptom individually; instead, we approached this dataset as a multi-label classification task with the purpose of achieving the ability to predict all symptoms at once. In addition, we trained and tested different models (pre-trained transformers methods) given our research hypothesis question. And we haven't attempted to predict depression from posts because as DSM-5-TR explains, 5 of these 7 depression symptoms must appear and cause clinically significant distress or impairment in social, occupational, or other important areas to be able to diagnose depression. Also, it should be noted that some of these symptoms are also present in other conditions; hence, without previous knowledge of the clinical record of the users, it would be very difficult to attribute these symptoms to a depressive episode (American Psychiatric Association, 2022). Although, Duvvuri et al. (2022),as part of their study, predicted depression they did that by making the strong assumption that these posts were written by truly depressed people. We however, do not share this assumption, as clinically significant distress or impairment cannot be assessed from only these posts, and additional data and medical records would be required. We thus, only take these predictions to be weak indications of users' emotional states that would need further investigation to provide a diagnosis.

Another study, used a pre-trained Whole Word Masking BERT model for multi-task learning to identify depressive symptoms as the primary task and figurative usage detection as the auxiliary task [6] (Yadav et al. (2020). They detected depression symptoms from tweets which were manually annotated based on the 9 categories of symptoms from PHQ-9 (Patient Health Questionnaire-9), and these tweets were also tagged with the figurative classes as metaphor and sarcasm. They demonstrated that including information about figurative language (FL) significantly improves the BERT model's robustness and reliability for distinguishing the depression symptoms. Hence, this article supports the argument that more domain specific transformers will perform better than BERT based models.

However, this research is also limited to self-reported diagnosis of depression which as the paper recognized is unreliable and cannot be reflective of the population of clinically depressed people. (Duvvuri et al., 2022). They also assume that depressive users often tend to use FL elements to describe their symptoms Duvvuri et al. (2022), given that assumption, they only labelled tweets of self-reported depressive users, but they do not provide any evidence or argument to support this assumption.

Other NLP methods have been used to detect depression symptoms.Yazdavar et al. (2017) trained an LDA model with a predefined set of seed terms to identify depression symptoms in tweets. Mukhiya et al. (2020) proposed a word-embedding model that incorporates contextually-diverse em-

bedding that, in turn, combines depression lexicons as well as emotional knowledge by using online forums texts. On the other hand, Ahmed et al. (2021) proposed a bidirectional Long Short-Term Memory (LSTM) architecture with an attention mechanism to extract depression symptoms on online forums. And Uddin et al. (2022) detected depression symptoms from public posts on a Norwegian information website with an LSTM-based RNN model.

Zhang et al. (2022) also identified that around 17% of the NLP research in mental illness detection have been done using transformer-based methods. Regarding the use of pre-trained transformers on depression detection in Social Media, transformers like Roberta, Bert, DistilBERT, Xlnet, and Electra have showed very high performance in a wide variety of evaluation metric in contrast with other traditional machine learning modes as SVM, bagging models, and linear classifiers (e.g., Malviya et al. (2021); Zeberga et al. (2022)); or even against other deep learning techniques like CNN, MLP, LSTM, and Bidirectional LSTM (e.g., Kabir et al. (2023); Zeberga et al. (2022)). In fact, Zeberga et al. proposed a BERT-Bi-LSTM model that improves the accuracy of text depression and anxiety classification from posts on Reddit and Twitter. With respect to symptoms diagnosis, there is almost no research on applying transformers to this task; however, Yadav et al. (2020) demonstrate that a multi-task Whole Word Masking BERT learning framework significantly improves the performance of the BERT model. As we can observe, a lot of different NLP models have been tried to diagnose symptoms of depression but it seems like domain specific transformers model have not been tested before and compared to general-use models before.

3 Methods

3.1 Architecture

We used language models with transformers, and fine-tuned selected models that are based on BERT and RoBERTa architectures. The BERT (Bidirectional Encoder Representations from Transformers) model, developed by Devlin et al. (2019), is a groundbreaking development in the field of NLP. RoBERTa (Robustly optimized BERT approach) is a state-of-the-art NLP model by Liu et al. (2019) that builds upon the BERT architecture.

3.1.1 Transformer architecture

The Transformer architecture developed by Vaswani et al. (2017) forms the foundation of BERT. Transformers process and comprehend the relationships between words in a phrase using a mechanism known as self-attention. The Transformer is made up of an encoder-decoder structure in which the encoder analyses the input sequence and the decoder produces the output sequence. The encoder is composed of a stack of identical layers. Each layer has two primary components: (1) Multi-head self-attention mechanism. (2) Feed-forward neural network.The decoder operates similarly to the encoder, but it also has an additional attention mechanism that extracts pertinent data from the encoders' created encodings (Vaswani et al., 2017; Alammar, 2018). Fig.1 shows the architecture of the Transformer Jia (2019).

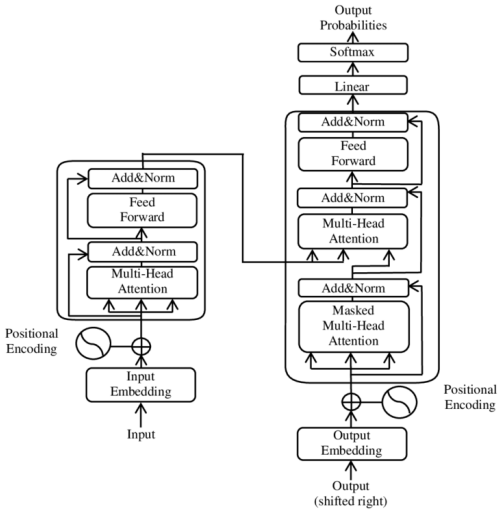

Figure 1: The Transformer model architecture

3.1.2 Pre-training and Fine-tuning

Pre-training is what enables BERT and RoBERTa to acquire general language comprehension from significant amounts of unsupervised data, while fine-tuning enables the models to be adapted to particular tasks using smaller amounts of labelled data. The effectiveness of BERT and RoBERTa in many NLP tasks depends on this two-stage training procedure.

Pre-training Tasks BERT is pre-trained using two unsupervised objectives; the Masked Language Model (MLM) task and the Next Sentence Prediction(NSP) task. RoBERTa, like BERT, uses MLM for pre-training but omits NSP.

- **Masked Language Model (MLM)** An MLM randomly masks some of the input tokens, and the model is trained to predict the original tokens based on the context that the surrounding unmasked tokens give (usually 15% of the tokens were masked) Clark et al. (2020). Due to the need that the model use both the words immediately before and immediately after the masked words to predict them, this training method enables BERT and RoBERTa to learn bidirectional representations from text.
- **Next Sentence Prediction (NSP)** NSP is a binary classification loss for determining whether two segments in the original text follow one another. It aids BERT in learning the contextual relationships between phrases.

Fine-tuning The pre-trained BERT model is adjusted to a particular NLP task during the fine-tuning stage. The labelled data is used to update the pre-trained model's weights during fine-tuning Quinn et al. (2020). By using this approach, BERT is able to apply the general language knowledge it acquired during pre-training to the precise needs of the target task. In terms of computing cost, Devlin et al. (2019) shows that pre-training is substantially more expensive than fine-tuning.

3.2 BERT, RoBERTa and BERT variants

The BERT-base model was used for this project, which comprises 12 Transformer blocks with 768 hidden units and 12 bidirectional self-attention heads and a total of 110 million parameters Devlin et al. (2019).Likewise,RoBERTa-base was used as it consists of 12 layers (transformer blocks), 12 attention heads, a hidden size of 768 units, and a total of 125 million parameters.

Regarding domain-specific models, two BERT variants were used : BERTweet and PHS-BERT.

BERTweet, a language model developed by (Nguyen et al., 2020), which has the same architecture as BERT and is trained using the RoBERTa pre-training process, has demonstrated to be highly effective in language processing. As have been trained on a large corpus of 850 million English tweets, BERTweet outperformed earlier state-of-the-art models when tackling NLP tasks on social media.

Naseem et al. (2022) proposed PHS-BERT for public health surveillance (PHS) on social media.With the same architecture as BERT (mask language modeling and next sentence prediction), PHS-BERT follows the BERT standard pre-training procedure, and it was initialized with the same weights from BERT in the training phase but trained on a corpus of health-related tweets. Naseem et al. (2022) also demonstrated that PHS-BERT achieved a robust performance and outperformed other transformers on social media depression classification tasks.

3.3 Exploratory Data Analysis (EDA)

3.3.1 Dataset

The dataset that was used was collected from Discord public servers' chat rooms tagged with "depression"; from there, they downloaded all posts and comments published from 20-04-2021 to 20-05-2021 (Duvvuri et al. (2022)).

For the labeling, a crowd sourcing technique was used where a group of 10+ Northeastern students from geographically diverse backgrounds, labeled "Y" if the symptom is identified in the post or "N" otherwise. The seven symptoms of depression that were labeled were defined by the American Psychiatric Association (APA) in DSM-5-TR, which is the principal authoritative handbook used by health care professionals around the world to diagnose mental disorders.

To understand these data we conducted data analysis on the dataset as seen below.

3.3.2 Data Overview

Dataset's structure The original data contained 23994 samples with the following attributes: **ID**, **AuthorID**, **Author**, **Date**, **Content of chat**, **Word Count** and **Seven indicators of target depressive symptoms** and **Non-depression**.The seven indicators of target depressive symptoms are:

1.Change in appetite, losing or gaining weight
2.Sleeping too much or not sleeping well (insomnia)
3.Fatigue and low energy most days
4.Feeling worthless, guilty, and hopeless
5.An inability to focus and concentrate that may interfere with daily tasks at home, work, or school
6.Movements that are unusually slow or agitated (a change which is often noticeable to others)
7.Thinking about death and dying; suicidal ideation or suicide attempts

3.3.3 Data Cleaning and Preprocessing

Cleaning and pre-processing text In the Content column all usernames were removed.The 'contractions' package was also used for the removal of stopwords after tokenization, and the 'WordNetLemmatizer' package was used for lemmatization. Both packages are from the 'nltk' library.

Column and row manipulation Columns with symptom labels were edited such that "N" was replaced with 0 and "Y" with 1. Rows with missing values were deleted as we made sure they were less than 0.25% and missing completely at random (MCAR)(Fig.2). The **ID**, **AuthorID**, **Author**, **Date**, **Word Count** and **Non-depression** columns were removed from the dataset. The Non-depression column was removed as all 0 symptom columns indicate no symptoms. The rest of the columns were irrelevant.

```
Number of columns with missing values: 8

Number of missing values in each column:
ID                                                                                    0
AuthorID                                                                              0
Author                                                                                0
Date                                                                                  0
Content                                                                               0
Words                                                                                 0
Change in appetite, losing or gaining weight                                         12
Sleeping too much or not sleeping well (insomnia)                                     7
Fatigue and low energy most days                                                      5
Feeling worthless, guilty, and hopeless                                               6
An inability to focus and concentrate that may interfere with daily tasks at home, work, or school    5
Movements that are unusually slow or agitated (a change which is often noticeable to others)          6
Thinking about death and dying; suicidal ideation or suicide attempts                 5
None                                                                                  5
dtype: int64

Total number of missing values in all columns: 51
```

Figure 2: Distribution of missing values

Class imbalances As we would like to make predictions based across all symptom columns we consider our target variable to be a combination of the values of the symptom label columns, and a class would look like this: [0, 1, 1, 0, 0, 1, 1, 1]. In this case, there are 98 different classes that are highly imbalanced with the most common occurrence being [0,0,0,0,0,0,0,1] at 70.67% (over 16000 cases) which means no symptoms and the second most common is [0,1,1,1,1,0,0,0] at 10.69% (Fig.3). To solve this, we have adopted a two-pronged approach. Firstly, we have computed the frequency of each class within the dataset and calculated their mean value. Classes with a frequency below the mean are oversampled by duplicating the data, which is then subjected to text augmentation via the "nlpaug" library. This technique involves replacing words in the text with their synonyms in a randomized fashion, effectively reducing the risk of overfitting. On the other hand, classes with a frequency above the mean are undersampled to maintain a proportional representation of all classes within the dataset.

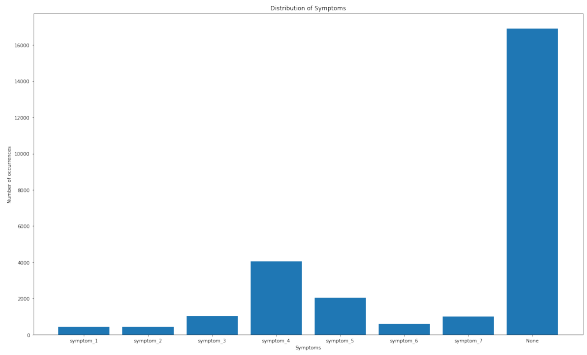

Figure 3: Distribution of symptoms

3.3.4 Exploratory Data Analysis

Fig.4 demonstrates that the text lengths in each labelled user comment range from 0 to 150 (ignoring long tail). As a result, the comments are not that long but it still gives us enough data on the topic of depression to be relevant.

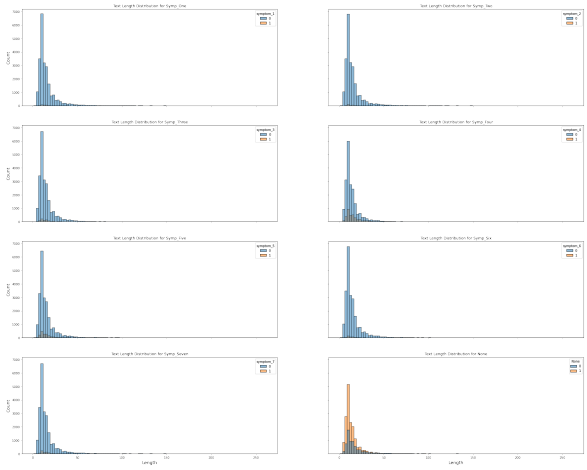

Figure 4: Distribution of text length for each symptom

Fig.5 shows a correlation matrix between symptoms.None of the symptoms seem to have a strong positive correlation with each other as none of the scores are above 0.6(generally considered to indicate moderate correlation). Therefore, since only weak correlations are present there seems to be a low danger for overfitting in respect to symptom correlation.

Barplots showing Top 10 Words and Top 10 Bigrams before and after preprocessing the dataset and Distribution of text length in user messages can be seen in Appendix A.

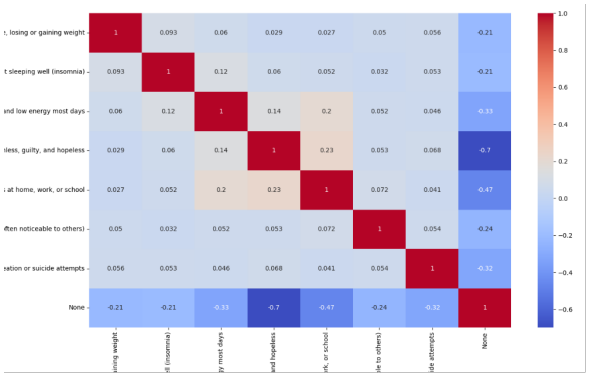

Figure 5: Correlation between symptoms

4 Experiments

4.1 Code

In our code, we started by initialising the tokenizers for each of the models that we used. The primary function of the tokenization process involves breaking down the input text into smaller units, which are then assigned unique numerical values. The tokenization of the input text results in a numerical representation of the text, which the machine learning model can process and make predictions on. The tokenizer for RoBERTa, known as RobertaTokenizer, and the one for BERT, called BertTokenizer, are utilised for this purpose. After that, in order to help the model better understand which parts of the input sequence the model focuses on for the task at hand, we create attention masks. These binary tensors are of the same shape as the input IDs tensor and serve to indicate which tokens should be attended to (marked as 1) and which should be ignored (marked as 0). The attention masks are then concatenated with the input IDs and fed into the model during training and inference. One of the key benefits of attention masks is that they allow the model to selectively focus on specific parts of the input sequence while ignoring others, particularly when dealing with inputs of varying lengths. This selective focus can improve both the accuracy and efficiency of the model, while also reducing the amount of irrelevant tokens that can introduce noise to the input. Once the data has been prepared for input, we divide it into training and testing sets. In our case, 20% of the data is reserved for testing while the remaining 80% is used for training. After dividing the data, the model is compiled using the Adam optimizer, which is a widely used optimization algorithm for deep learning models. The binary cross-entropy loss function is chosen, as it is commonly used for binary classification problems. Additionally, the binary accuracy metric is used to evaluate the model's performance during training and testing. Initially all models were trained on 10 epochs, the models were then trained on the training set with the number of epochs chosen based on the specific model that was being used, to ensure that we avoid overfitting. A batch size of 16 was used, which means that the model will update its weights after every 16 samples have been processed.

4.1.1 Analysis

After training the model on the training set, it's important to evaluate its performance on the test set to determine how well it can generalise to new, unseen data and to do some analysis using several techniques so that we can make sure the model is behaving as expected and to test its performance. Important evaluation metrics like precision, recall, and F1 score were calculated using the classification report from the 'sklearn' library. In addition, several metrics like micro-averaged and macro-averaged precision, recall, and F1 score, Hamming Loss, and Jaccard Score are computed to provide insights into the model's strengths and weaknesses. Confusion matrices are useful to evaluate the model's performance, as they show how many predicted labels are correct (true positives), incorrect (false positives), and missed (false negatives). Normalised confusion matrices provide a more detailed performance analysis of the classification model. The F1 score is a widely used metric to evaluate the performance of a binary classification model. It combines both precision and recall and is useful when classes are imbalanced. Exact match ratio and all-correct percentage metrics measure the percentage of instances where all predicted labels match the true labels. Examining training and validation loss over multiple epochs can help visualise trends and determine whether the model is overfitting or underfitting. Error analysis is an important step to identify patterns and areas of improvement by analysing the errors made by the model on the test data. Additionally, for each model, the attention weights are analysed by computing the average attention weights for each token pair across all layers and then selecting the top heads based on the average weights for the layers 1 and 12. The first and last layers were chosen as they can provide insights into the initial and the final refined representation of the input. These top

heads are then used to plot a heatmap that visualizes their attention weights. Heatmaps were plotted for two sentences: 1) "Can't sleep, always tired, hard to focus." 2) "Lost appetite, feel guilty, thinking of ending it.". These sentences cover symptoms 2, 3, and 5 (insomnia, fatigue, and inability to focus) and 1, 4, and 7 (change in appetite, feelings of worthlessness, and suicidal ideation) respectively. This information can be used to refine the model and improve its performance on unseen data.

5 Results and Discussion

5.1 Examining training and validation loss upon initial training

As previously mentioned initially all models were trained on 16 batch size and 10 epochs with early stopping to understand when they are overfitting the data. For this the training and validation loss were plotted for each model. For all models, training and validation loss initially decrease significantly, however all models seemed to reach an epoch number where models became less generalizable and started to overfit the data. Specifically, that number of epochs was 2 for BERT-base, 2 for RoBERTa-base, 4 for BERTweet and 1 for PHS-BERT (Fig.6). The models were thus, later trained on the amount of epochs specific to them to prevent overfitting (Appendix B Fig.10).

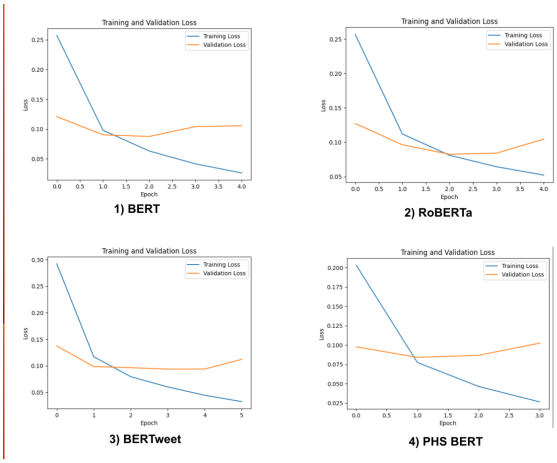

Figure 6: Training and Validation loss plot for models

5.2 Accuracy, Precision, Recall, F1-score and EMR

Table 1 displays that BERT showed the highest test accuracy (0.97) and lowest test loss (0.09). It demonstrated strong performance across all symptoms with consistently high precision, recall, and F1-scores. RoBERTa and BERTweet have the similar performance, as their precision, recall, and F1-scores were slightly lower than BERT. PHS-BERT had a test accuracy of 0.97 and test loss of 0.09, performing almost as well as BERT. It had high precision, recall, and F1-scores, but slightly lower than BERT. Figure shows the metrics for each symptom individually. More detailed classification reports (Appendix B Fig.11).

Model	Test Accuracy	Test Loss
BERT	0.97	0.09
RoBERTa	0.72	0.71
BERTweet	0.72	0.69
PHS-BERT	0.97	0.09

Table 1: Test accuracy and test loss for BERT, RoBERTa, BERTweet, and PHS-BERT.

Additionally, the Exact Match Ratio (EMR) was obtained which represents the ratio of instances with perfect matches between the true and predicted labels across all labels. For BERT this was 85%, for RoBERTa 83%,for BERTTweet 85% and for PHS-BERT 84%.

5.3 Hamming Loss (HL) & Jaccard Score (JS)

The BERT model had a HL of 0.027 and a JS of 0.53. RoBERTa-base model had a score of 0.03 and 0.52 for HL and JS respectively. The BERTweet model had HL of 0.0266 and a JS of 0.5357. Finally, the PHS-BERT model yielded a Hamming Loss of 0.0306 and a Jaccard Score of 0.5168. The BERTweet thus showed the best performance among the four models, with the lowest HL and the highest JS. Meanwhile, the PHS-BERT model exhibited the weakest performance, with the highest HL and the lowest JS.

5.4 Confusion Matrices

A confusion matrix is a valuable tool for assessing the performance of a classification model, particularly in binary classification problems. It is a table that summarises the results of a classification algorithm by displaying the number of true positives, false positives, false negatives, and true negatives. In order to gain a deeper understanding of our results, we decided to create a confusion matrix for each of the 7 symptoms. One key observation is that all four models produced comparable results. This indicates that the models were trained effectively and that there were no significant differences

in their ability to classify symptoms. However, upon closer inspection, we can see that the BERT model outperformed the other models, with slightly lower numbers of false positives and false negatives. Another important finding is that all models struggled the most with the fourth symptom. This suggests that there may be specific challenges associated with this symptom that the models were not able to fully address. Further investigation is needed to identify the reasons for this and to improve the models' performance on this symptom.

5.5 Error analysis

Error analysis is a crucial step in improving the performance of a machine learning model. By examining the data that was incorrectly classified by the model, we can identify patterns or trends that may have caused the errors. In our case, we believe that some errors were due to the manually classified dataset which could have contained some misclassifications. However, this does not discount the importance of error analysis as it provides valuable insights into the strengths and weaknesses of our models. It is interesting to note that out of the models tested, BERT and BERTweet had the lowest number of errors with an exact match of 85%. This suggests that these models have a higher level of accuracy and may be better suited for this particular task.

5.6 Attention weight visualisation

The 5 top Layer Head combinations were analysed below for each sentence for BERT(Appendix B Fig.12), RoBERTa (Appendix B Fig.13), BERTweet (Appendix B Fig.14) and PHS-BERT (Appendix B Fig.15).

5.6.1 Sentence; "Can't sleep, always tired, hard to focus."

In BERT, most Layer 1 heads show high attention to the [CLS] token (0.75), with L1H4 and L1H3 focusing on the negation "can't" with attention values of 0.7 and 0.85, respectively. Attention weights vary for other tokens, revealing the model attends to different words and relationships. For L12H12, less attention is paid to [CLS] and [SEP], while more is paid to key concepts like "tired" (0.3 to 0.6), "sleep" (0.23 and 0.5), and "focus" (0.3 to 0.4), highlighting their importance. For RoBERTa, L1H4 and L12H12 emphasize key words, L1H3 (0.75) highlights self-attention on the central issue, and L12H11 connects context with sleep and tiredness. The matrices showcase the model's understanding of the main problem through various attention patterns. In BERTTweet, L1H4 has high values concentrated along the diagonal, suggesting token-level feature extraction. L1H3 captures a mix of token-level features and relationships, with some tokens having high attention values (e.g., "Can" at 0.26). L2H2 and L12H11 aggregate information into the CLS token, while L12H12 attends to specific tokens and sentence structure for context-aware representations (e.g., "hard" has high attention to "Can" at 0.2).For PHS-BERT, L1H4, L1H3, and L1H2 have high diagonal values and focus on key phrases like "Can't sleep" (0.3), "always tired" (0.4), and relationships such as "always tired" and "hard to focus" (0.2). L12H15 and L12H16 both focus on the [CLS] token (0.9 and 0.8), with L12H15 emphasizing stronger connections between words like "tired," "hard" (0.3), "to" (0.4), and "focus" (0.2) compared to L12H16.

5.6.2 Sentence; "Lost appetite, feel guilty, thinking of ending it."

In BERT's Layer 1, the [CLS] token shows strong attention (0.75) with itself in L1H4, L1H3, and L1H11, focusing on overall sentiment. For L1H4, attention values between [CLS] and "lost" (0.5) and "appetite" (0.25) indicate context understanding. In L12H12, off-diagonal attention values suggest relationship understanding. RoBERTa's L1H4 has evenly distributed attention, while L1H3 captures overall context, and L1H2 focuses on the initial token. L12H11 and L12H12 emphasize the first and second-to-last tokens. BERTTweet's L1H4 and L1H3 attend to multiple tokens, with selective focus on different input parts based on context. In L1H3, concentration suggests model identification of specific relevant tokens. L12H11 and L12H12 have concentrated diagonal values, differing in off-diagonal attentions. For PHS, L1H4 exhibits dispersed attention values, with the highest in row 10 (0.5). L1H3 focuses on the first token (0.9), while L1H2 presents an even distribution, with a peak value of 0.7 in the first row. L12H15 has high attention on the first token (0.9), and L12H16 displays more even distribution, with the highest value of 0.7 in row 3.

5.7 Discussion

When comparing accuracy, precision, recall, and F1-scores, BERT-base outperformed the other models, closely followed by PHS-BERT. Both

RoBERTa-base and BERTweet showed similar performance but lagged behind BERT-base and PHS-BERT. In terms of the Exact Match Ratio, BERT and BERTweet had the highest scores, suggesting better accuracy and suitability for the task. Hamming Loss and Jaccard Score results revealed that BERTweet had the best performance, with the lowest HL and the highest JS, while PHS-BERT had the weakest performance. Confusion matrices indicated that all models produced comparable results, with BERT-base slightly outperforming others. It is worth noting that all models struggled with the fourth symptom, which may require further investigation to identify potential challenges and improve model performance. When comparing the attention patterns of BERT, RoBERTa, BERTTweet, and PHS-BERT, it's clear that each model attends to different aspects of the input sentence and exhibits varying strategies to understand and represent the text.Specifically,BERT's attention is concentrated on the [CLS] token and negation (can't) in early layers, indicating sentence-level understanding and the importance of negation. RoBERTa emphasizes keywords and highlights self-attention on the central issue, connecting context with sleep and tiredness. BERTTweet exhibits high attention values along the diagonal in some heads, implying a focus on token-level features and relationships between specific tokens, while also aggregating information with the [CLS] token. PHS-BERT displays high diagonal values in early layers and highlights key phrases such as "Can't sleep" and "always tired," while also capturing relationships between tokens like "tired" and "hard to focus" in later layers.

5.8 Limitations and future work

A limitation of our dataset is the lack of supervision by medical experts in the assessment and annotation of depression symptoms; therefore, our research is based on a labelled dataset that might be biased or prone to misclassification. Another possible shortcoming of our dataset is that it may incur on sample biases by relying on users to join a forum. In addition, it is necessary to collect and label more data and from other social media sites to further replicate and confirm our results. Our sample is possibly not diverse enough to be able to identify all the possible combinations and manifestations of depression symptoms that are often present in the population.

Regarding our approach, we only focused on the two of the best-known general pre-trained transformers and the two more domain-specific BERT models, but perhaps other transformers may have better performance. Thus, future research should be focused on including other pre-trained methods to broaden the benchmark spectrum of models for symptom depression classification.

In this research, we worked with de-identified data, so we could not follow user's posts over time as well as the conversation between users to have a better accuracy on depression symptoms detection; then, a potential future research is to design an experiment in which it is possible to follow posts over time and their interactions with other users.

Finally, integrating our research approach with clinical data might be enough to be able to identify depression and not just depression symptoms since APA clinical guidelines to diagnose depression require at least 5 symptoms and a clinically significant distress or impairment caused by these symptoms.

6 Conclusion

In conclusion, our study provides some insight to the performance of BERT, RoBERTa, BERTTweet, and PHS-BERT in symptom identification. BERT-base performed the best, followed by PHS-BERT, with BERTTweet showing optimization potential. Attention analysis exposed unique model strategies in processing input text. Despite promising outcomes, addressing limitations and exploring future work will improve performance and applicability in clinical settings, aiding healthcare professionals in providing accurate and timely patient care.

7 Ethics

No ethics approval was necessary for the dataset analysed in this study because this data was drawn from a published study, and it was completely de-identified. However, the further use of such techniques in real world scenarios would require thorough ethical approval.

References

Usman Ahmed, Suresh Kumar Mukhiya, Gautam Srivastava, Yngve Lamo, and Jerry Chun-Wei Lin. 2021. Attention-based deep entropy active learning using lexical algorithm for mental health treatment. *Frontiers in Psychology*, 12.

Jay Alammar. 2018. The illustrated transformer.

American Psychiatric Association. 2022. *Diagnostic and statistical manual of mental disorders, Text Revision (DSM-5-TR)*, 5th ed. edition. Autor, Washington, DC.

Silvia Casola, Ivano Lauriola, and Alberto Lavelli. 2022. Pre-trained transformers: an empirical comparison. *Machine Learning with Applications*, 9:100334.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators.

Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cécile Paris. 2020. Cost-effective selection of pretraining data: A case study of pretraining BERT on social media. *CoRR*, abs/2010.01150.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Venkata Duvvuri, Qihan Guan, Swetha Daddala, Mitch Harris, and Sudhakar Kaushik. 2022. Predicting depression symptoms from discord chat messaging using ai medical chatbots. In *2022 The 6th International Conference on Machine Learning and Soft Computing*, ICMLSC 2022, page 111–119, New York, NY, USA. Association for Computing Machinery.

Yuening Jia. 2019. Attention mechanism in machine translation. *Journal of Physics: Conference Series*, 1314(1):012186.

Mohsinul Kabir, Tasnim Ahmed, Md. Bakhtiar Hasan, Md Tahmid Rahman Laskar, Tarun Kumar Joarder, Hasan Mahmud, and Kamrul Hasan. 2023. Deptweet: A typology for social media texts to detect depression severities. *Computers in Human Behavior*, 139:107503.

Anant Kumar and K. Rajasekharan Nayar. 2021. Covid 19 and its mental health consequences. *Journal of Mental Health*, 30(1):1–2. PMID: 32339041.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Keshu Malviya, Bholanath Roy, and SK Saritha. 2021. A transformers approach to detect depression in social media. *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pages 718–723.

Suresh Kumar Mukhiya, Usman Ahmed, Fazle Rabbi, Ka I Pun, and Yngve Lamo. 2020. Adaptation of idpt system based on patient-authored text data using nlp. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 226–232.

Usman Naseem, Byoung Chan Lee, Matloob Khushi, Jinman Kim, and Adam Dunn. 2022. Benchmarking for public health surveillance tasks on social media with a domain-specific pretrained language model. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 22–31, Dublin, Ireland. Association for Computational Linguistics.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets.

Office for National Statistics. 2021. Coronavirus and depression in adults, great britain: July to august 2021. Technical report.

Joanne Quinn, Joanne McEachen, Michael Fullan, Mag Gardner, and Max Drummy. 2020. *Dive into deep learning: Tools for engagement*. Corwin, a SAGE Company.

Md. Zia Uddin, Kim Dysthe, Asbjørn Følstad, and Petter Brandtzaeg. 2022. Deep learning for prediction of depressive symptoms in a large textual dataset. *Neural Computing and Applications*, 34:1–24.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

World Health Organization. 2022a. Mental disorders. Technical report.

World Health Organization. 2022b. Mental health and covid-19: Early evidence of the pandemic's impact. Technical report, World Health Organization.

Shweta Yadav, Jainish Chauhan, Joy Prakash Sain, Krishnaprasad Thirunarayan, Amit Sheth, and Jeremiah Schumm. 2020. Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework.

Amir Yazdavar, Hussein Al-Olimat, Monireh Ebrahimi, Goonmeet Kaur Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. 2017. Semi-supervised approach to monitoring clinical depressive symptoms in social media. volume 2017.

Kamil Zeberga, Muhammad Attique, Babar Shah, Farman Ali, Yalew Zelalem Jembre, and Tae-Sun Chung. 2022. A novel text mining approach for mental health prediction using bi-lstm and bert model. *Computational Intelligence and Neuroscience*, 2022.

Tianlin Zhang, Annika Schoene, Shaoxiong Ji, and Sophia Ananiadou. 2022. Natural language processing applied to mental illness detection: A narrative review. *npj Digital Medicine*.

A Appendix

The left plot from Fig.7 displays the top 10 words together with the number of times each word appears before preprocessing. The top words from the original dataset are 'to, and, I, you, the' with a high frequency, as can be seen in the figure. The word 'to' has a frequency close to 26000 followed by 'and ' and 'I'. The right plot from Fig.7 shows the Top 10 Most Frequently Occurring Bigrams before preprocessing. It can be seen that 'if you', 'want to', 'to be' are the three most frequent bigrams, which with a high frequency around 2000. These high frequency words and bigrams are all essentially stop words and will be removed in the pre-processing stage to obtain more valid data.

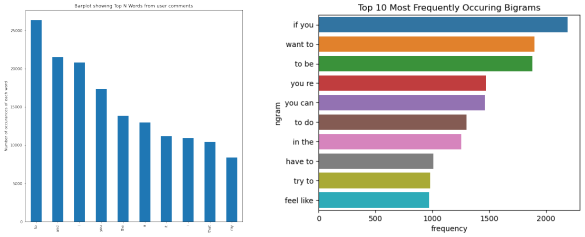

Figure 7: Barplots showing Top 10 Words and Top 10 Bigrams

Fig.8 shows barplots for Top 10 Words and Top 10 Bigrams after preprocessing.

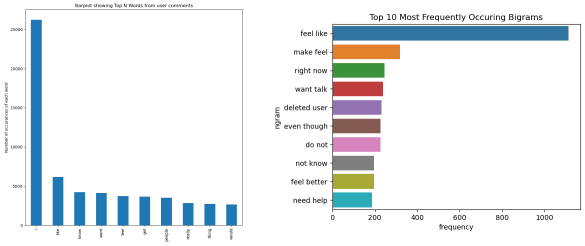

Figure 8: Barplots showing Top 10 Words and Top 10 Bigrams after preprocessing

Fig.9 shows the distribution of text length in user messages.

B Appendix

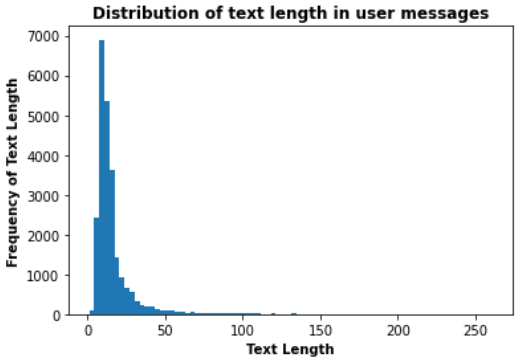

Figure 9: Distribution of text length in user messages

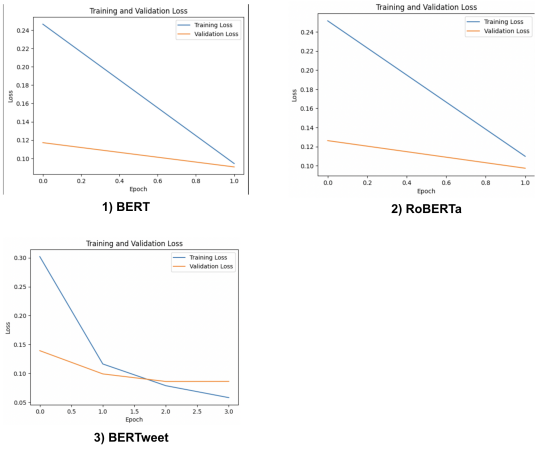

Figure 10: Train loss and validation loss plot after selection of epochs

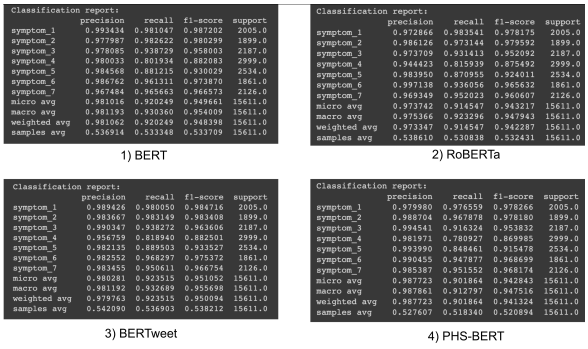

Figure 11: Classification reports for the 4 models

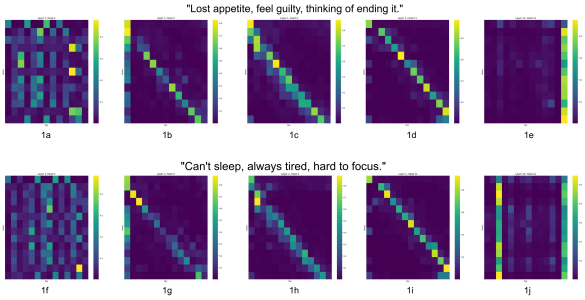

Figure 12: BERT attention weights visualization

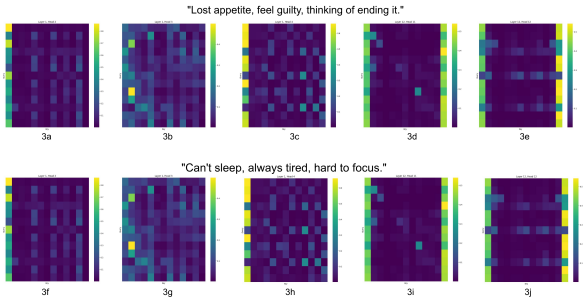

Figure 13: RoBERTa attention weights visualization

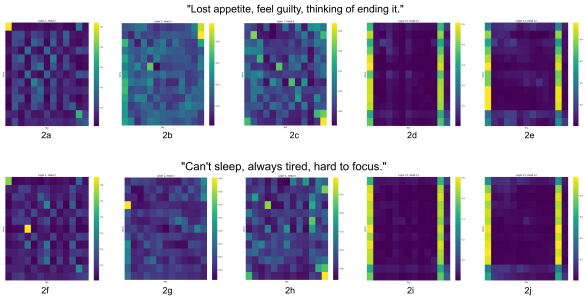

Figure 14: BERTTweet attention weights visualization

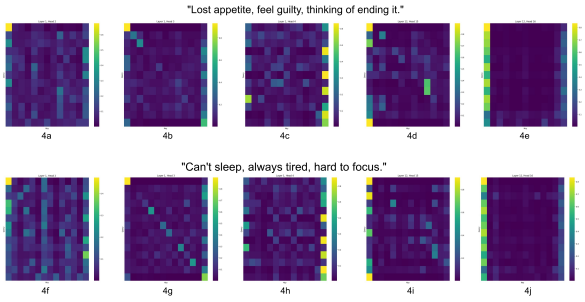

Figure 15: PHS-BERT attention weights visualization