ISSN 1816-6075 (Print), 1818-0523 (Online)
Journal of System and Management Sciences
Vol. 14 (2024) No. 3, pp. 318-336
DOI:10.33168/JSMS.2024.0318

A Comparative Study of Deep Learning Models for Detecting Depressive Disorder in Tweets

Natasha Alyaa Anindyaputri, Abba Suganda Girsang

Computer Science Department, Binus Graduate Program – Master of Computer Science
Jakarta, Indonesia
natasha.anindyaputri@binus.ac.id; agirsang@binus.edu

Abstract. Depressive disorder is a psychiatric problem that has a high contribution to the cause of disability in the world, as it can cause people who experience it to not perform well in daily life which might affect their work life. To help patients who needed immediate treatment, the World Health Organization (WHO) has created a program called Mental Health Gap Action Program (mhGAP) in order to increase development that focuses on monitoring mental health in a globalized world. This research explores several different classification model architectures with deep learning approaches in order to detect Twitter users' tweets that were classified as tweets with depressive disorder and then compare the results between the proposed models. As a result, the Long Short-Term Memory (LSTM) model paired with Transformer word representation models has provided better performance when compared to employing Bidirectional Encoder Representations from Transformers (BERT) baseline classification or LSTM classification models alone. The evaluation of the classification model's architecture reveals that the LSTM model, when paired with MentalBERT, exhibits the highest accuracy performance, achieving an accuracy score of 0.86. This approach leverages the respective advantages of transformer and LSTM models to facilitate the accurate analysis of syntactic and contextual information associated with individual words, hence enabling the appropriate representation or domain-specific of sequential data.

Keywords: Depressive disorder, Twitter, Deep Learning, LSTM, BERT, Transformer.

1. Introduction

According to the data statistics from the Global Health Metrics, depressive disorder is a psychiatric problem that has a high contribution to the cause of disability in the world. Depressive disorder is one of the psychiatric problems that are taken very seriously (WHO 2020), because it can cause the people who experience it to not perform well in daily life which might affect their work, school, or even the way they interact with their family or loved ones. The main characteristics of people who experience depressive disorders are anhedonia (difficulty in finding pleasure) and a mood inclined to depression. If these characteristics occur, then the next criteria are examined. There might be some somatic complaints, such as sleep problems, drastically decreasing appetite and weight, difficulty in concentrating, experiencing mental fatigue, and a decreased ability to respond well to emotions (irritability, tiredness, or anxiety). Some people might also experience some non-somatic disorders, such as feeling sad or helpless, losing the ability to feel pleasure, feeling worthless, feeling guilty, or having suicidal thoughts.

Recent statistics underscore the staggering scale of depressive disorders on a global scale. According to the World Health Organization (WHO), depression is one of the leading causes of disability worldwide, affecting more than 264 million people as of the most recent data. Moreover, the COVID-19 pandemic has further exacerbated the situation, with a surge in reported cases of depression and anxiety due to the social, economic, and health-related challenges it has presented. A recent investigation published in The Lancet (Santomauro et al. 2021) reveals that the COVID-19 pandemic has resulted in a significant increase in global rates of depressive and anxiety disorders as a result of the COVID-19 pandemic in 2020. The study revealed a substantial surge in the overall prevalence of mental disorders, with an additional 53.2 million cases of anxiety disorders and 76.2 million cases of major depressive disorders (MDD). This data shows a significant increase from prior years, illustrating the rising concern over mental health.

Based on the statistical data from the World Health Organization (WHO), at least 1 out of 4 people in the world will be affected by mental problems at some point in their lives (Pavlova and Berkers 2020). Some studies have found that mental health problems can result in a suicide attempt (Sarsam et al. 2021). Pavlova & Berkers (Pavlova and Berkers 2020) cited that according to the Disability Adjusted Life Years (DALYs), there are approximately 450 million people in the world with mental health problems, out of which 62% of them having a tendency to attempt suicide. Despite the fact these mental health problems can be treated with some medication, according to a statement from Chisholm, that was cited by Pavlova & Berkers (2020), the treatment of psychiatric problems is usually hampered due to the patients not realizing that their mental disorders are already at a level that needs professional help immediately (Asad et al. 2019), and the inadequate related policies for handling people with mental health problems. For these reasons, the WHO established a program called the WHO Mental Health Gap Action Program (mhGAP) related to the development of information systems that focus on monitoring mental health in a globalized world (Zhou, Hu, and Wang 2019). This has prompted many researchers to consider exploring the content of social media platforms in order to understand the viewpoint of users with psychiatric problems and what drives them to have some thoughts of committing suicide (Sarsam et al. 2021).

Twitter is now widely recommended and considered by many researchers for conducting research related to the mental and emotional condition of the user. It is considered one of the tools that have an important contribution for social scientists in collecting large amounts of data in a short time, with the capability to process text up to 280 characters (Reynard and Shirgaokar 2019). Researchers can analyze the sentiment expressed in tweets related to the relevant topics. However, due to the large number of tweets, it is not possible for researchers to analyze the data qualitatively. For instance, out of 24,500 tweets collected in relation to mental health problems, only 3% could be used for the analysis. Therefore, researchers currently use machine learning and deep learning related to natural language for analyzing the tweets of Twitter users (Reynard and Shirgaokar 2019).

Long Short Term Memory (LSTM) has been one of the most popular architectures and has emerged as an effective and scalable model for several learning problems related to sequential data (Greff et al. 2017). Due to its capacity to recognize long-term dependencies, LSTM is employed to advance the state-of-the-art solutions for many difficult problems, such as in classification systems, question answering, and text generation and/or prediction tasks (Martínez-Castaño et al. 2021). Winata et al. utilized LSTM models with word embeddings to classify any psychological stress from interview transcriptions. The models were trained to discover the statistical aspects of language used when under stress by learning the patterns of the words and phrases that are used more frequently when someone is stressed. However, the model did not pick up on more complicated grammatical structures. While LSTM models may be effective in identifying patterns in language usage, they may not be able to fully capture the complexity of grammatical structures to process the semantic meanings that could be caused by a lack of data.

One of the key limitations of existing methods is their reliance on traditional machine learning algorithms and static word embeddings. This presented numerous challenges, including the requirement for all meanings of a term with multiple senses to possess a unified representation, which could often struggle to capture the intricate nuances present in mental health narratives. The inherent challenge lies in the dynamic and context-dependent nature of language, where the same words can have different meanings depending on the surrounding context (Ethayarajh 2019). Furthermore, existing methods often lack the ability to comprehend complicated phrase structures and sentiment variations within narratives, making them less effective in identifying subtle emotional cues indicative of mental well-being.

A new trend in Natural Language Processing (NLP), contextualized pre-trained language models, has received a lot of interest for a variety of text processing applications (Ji et al. 2022), which is known as Transformer. Transformer has proven its ability to understand global contexts and offers a variety of modeling languages to support different word representations (Ethayarajh 2019; Miaschi and Dell'Orletta 2020). The recent advancements in context-dependent embeddings, as demonstrated by Peters et al. (2018) and Devlin et al. (2019), have proven to be highly effective in achieving state-of-the-art performance across various intricate NLP tasks. One of the most popular architectures of transformer is BERT (Martínez-Castaño et al. 2021), which stands for Bidirectional Encoder Representations from Transformers (Devlin et al. 2018). BERT has shown promising results in various NLP tasks, such as question answering, sentiment analysis, and text classification (Martínez-Castaño et al. 2021). BERT's success has led to the development of various domain-specific pre-trained language models for learning text representations, such as the biomedical BERT (Lee et al. 2020), clinical BERT (Alsentzer et al. 2019), and also mental healthcare related domains like MentalBERT (Ji et al. 2022). These domain-specific pre-trained language models have shown great promise in improving the accuracy and effectiveness of natural language processing tasks within their respective fields.

Although previous research using LSTM models has shown that they perform reasonably well in classifying mental health conditions, using domain-specific pre-trained language models may improve accuracy and efficiency by incorporating recent developments in contextual word representations like MentalBERT. This study will examine various alternative deep learning classification models, including the LSTM, BERT classification model, and models that combine the LSTM and Transformer word representation models. The combination of these two models make use of the benefits of LSTM in assessing sequence and temporal relationships in text, as well as the benefits of Transformer models in recognizing global contexts and better processing word representations.

2. Literature Review

2.1. Prior Work Applying Maching Learning and Deep Learning for Mental Health Classification

Much research had been carried out on the implementation of deep learning and machine learning for supporting developments related to psychiatric problems. The implementations of machine learning related to psychiatric problems require various supporting data, such as data of user behavior pattern, data of interactions, as well as data related to the emotions of users (Zhou, Hu, and Wang 2019).

Winata et al. (2018) evaluated various models to classify psychological stress from self-conducted interview transcriptions with distant supervision techniques, which enhanced the size of the used corpus by automatically classifying tweets based on their hashtag content. The model parameters were fine-tuned using the interview and the Twitter corpus data. Some of the models evaluated in this study are LSTM, Bidirectional Long Short-Term Memory (Bi-LSTM), and the addition of attention layer to LSTM and Bi-LSTM models. The addition of attention mechanisms in the conducted research enabled the model to choose instructive words. The best performance in terms of accuracy (74.1%) and F1-score (74.3%) was attained by applying the Bi-LSTM model with attention. However, the model did not pick up on more intricate grammatical structures. The observation indicated that the model did not pick up on the semantic significance of negation, which may be due to a dearth in the training data. Furthermore, Winata considered that the Twitter corpus might be a poor source for teaching models on how to use correct grammatical structures.

Zhou, Hu, and Wang (2019) developed a model referred to as Mental-Disorder-Aided-Diagnosis (MDAD) model to predict the number of Twitter users who experienced psychiatric problems such as depressive disorder, obsessive-compulsive disorder, bipolar disorder, anxiety disorder, and panic disorder, Zhou et al. (2019) also developed a sentiment analysis method for providing the emotional values of anger, loathing, disgust, grief, sadness, surprise, terror, fear, trust, joy, and anticipation. It used NRC-Lexicon as an initial corpus and added synonyms obtained from WordNet, then weighted based on the K-Means clustering algorithm. Then the Twitter users were categorized based on the probability value obtained from the MDAD model. If the probability value reaches more than 80%, then it will be categorized as a user with a severe mental disorder. For a probability value that is between 60% and 80%, it will be categorized as a user with a moderate mental disorder, while for a probability value below 60%, it will be categorized as a user with a mild or no mental disorder. To optimize the model that has been developed, the Stochastic Gradient Descent (SGD) algorithm was used, and the test was carried out 10 times for each condition of the psychiatric problem.

Kim et al., (2020) used deep learning to predict social media users with mental disorders such as autism, depression, anxiety, bipolarity, and schizophrenia based on Reddit social media posts published on subreddit channels r/mentalhealth, r/depression, r/Anxiety, r/bipolar, r/BPD, r/schizofrenia, and r/autism. After pre-processing, the data was obtained from 228,060 users, with a total of 488,472 posts. Kim et al. (Kim et al. 2020) employed the Synthetic Minority Over-Sampling approach to handle unbalanced data. The XGBoost and CNN algorithms were used to conduct binary classification for each category of existing mental diseases. On the XGBoost classification algorithm, Kim et al. (Kim et al. 2020) implemented TF-IDF, while on the CNN model, text data was processed using Word2Vec. As a result, classifications using the CNN model obtained higher accuracy values on each proposed classification model, 75.13% for depression, 77.81% for anxiety, 90.20% for bipolarity, 90.49% for BPD, 94.33% for schizophrenia, and 96.96% for autism. The classification with the XGBoost model obtained accuracy values of 71.69% for depression, 70.41% for anxiety, 85.53% for bipolarity, 85,14% for BPD, 86.72% for schizophrenia, and 94.91% for autism.

2.2. Background on LSTM Models

LSTM is one of the types of Recurrent Neural Network (RNN) architecture designed to solve the problem of vanishing gradients (Greff et al. 2017). LSTM has emerged as an effective model and can

be adapted to a variety of issues related to sequential data (Greff et al. 2017). LSTM architecture has been specifically developed to effectively retain and utilize long-term data information through the incorporation of input gate, output gate, and forget gate operations (Minaee et al. 2020). The LSTM model has demonstrated its efficacy in phrase classification by yielding precise outcomes, and it has been increasingly employed in many NLP endeavors (Zeberga et al. 2022).

Tadesse et al. (2020) proposes an automated identification system for detecting suicide posts on the Reddit social media platform. The system utilizes deep learning and machine learning-based categorization techniques. The researchers utilized a combined model consisting of LSTM-CNN architecture to assess its performance and conduct a comparative analysis with the other classification algorithms. The suggested model utilizes the output vector of the LSTM as the input vector of the Convolutional Neural Network (CNN). Subsequently, a novel CNN model is constructed upon the LSTM architecture in order to extract notable characteristics from the input textual phrases, enhancing the accuracy of the classification outcomes. The researchers utilize a Word2Vec technique to generate word embeddings for the LSTM and CNN models. The LSTM-CNN combined model with word embedding techniques achieved the best classification results for relevance, demonstrating the power and capability of deep learning architectures for suicide risk assessment, achieving 93.8% accuracy and an F1 score of 92.8%.

Lin et al. (2020) suggested a model for detecting depression using audio signals along with linguistic content obtained from the patient's interviews. This research implemented the Bi-LSTM method with the addition of attention layers to address the linguistic data used. To process the sound signals, Lin et al. (Lin et al. 2020) applied the one-dimensional convolutional network algorithm (1D CNN). Lin et al. (Lin et al. 2020) used pre-existing datasets, the Distress Analysis Interview Corpus-Wizard-of-Oz (DAIC-WoZ) and Audio-Visual Depressive Language Corpus (AViD-Corpus), as model initiatives. To measure the severity of depression, Lin et al. (2020) used the Patient Health Questionnaire (PHQ-8) and Beck Depression Index-II (BDI-II). The research conducted that the proposed method could assess the severity of depression efficiently.

Ghosh et al. (2021) presents a novel hybrid deep learning framework for forecasting the potential influence of the COVID-19 pandemic on mental well-being based on data extracted from Twitter social media platforms. This study examines more than 571,000 tweets posted by 482 users between October 2019 and May 2020. Notably, a substantial increase in the number of tweets expressing depressive sentiments was seen between the months of February and May in 2020. The LSTM-CNN neural network was utilized. The predictive performance of the model in detecting depression yielded an accuracy rate of 99.42% and a f1-score of 0.9916. Out of the total 3,104 testing data, the model accurately classified 2,401 tweets as non-depressed and 685 tweets as depressive.

Similar to Winata et al. (2018), Almars (2022) study performed a depression analysis on Arabic social media content using Bi-LSTM with an attention mechanism. The proposed deep learning model combined an attention mechanism with a Bi-LSTM to simultaneously focus on discriminative features and learned significant word weights that contribute highly to depression detection. The experiment results showed that the proposed model outperformed state-of-the-art machine learning models in detecting depression. The attention-based Bi-LSTM model achieved 83% accuracy on the depression detection task.

2.3. Evolution of Contextual Word Representation

Understanding natural language from text data is an essential area of Artificial Intelligence. In the same way that images and acoustic waves can be mathematically modeled by analog or digital signals, we need a method to represent text data to automatically process and understand the meaning of it. However, one of the problems in NLP development is that as more vocabulary is learned, the larger the dimensions of the features used (Li and Yang 2018). Therefore, having a way for representing existing vocabulary in smaller dimensions is crucial.

Traditionally, these word embeddings were static, which assigned a fixed vector representation to each word regardless of its context (Ethayarajh 2019; Mikolov et al. 2013). While these embeddings were valuable for capturing word semantics, they fell short in capturing nuances like polysemy and context-dependent meanings. Recent work, namely deep neural language models such as ELMo (Peters et al., 2018) and BERT (Devlin et al. 2018), has successfully constructed contextualized word representations, word vectors that are sensitive to their context of occurrence.

Miaschi and Dell'Orletta (2020) compares the linguistic knowledge encoded in the internal representations of a contextual Language Model (BERT) and a contextual-independent one (Word2vec). The development of contextual word representation has made it possible to encode sentence-level features even inside single word embeddings and to comprehend the complete context of each word in an input sequence.

By contextualizing word meanings depending on their surrounding context, these models go beyond conventional word embeddings and hold enormous promise for categorizing mental health conditions.

2.4. Applications of BERT for Mental Healthcare

BERT's contextual word representations have empowered models to achieve remarkable performance in sentiment analysis, emotion detection, and mental health disorder classification. It has played a pivotal role in understanding and extracting insights from patient narratives, online forums, and clinical records.

Syed and Narvekar's research (2021) demonstrated how BERT can be used to support the detection of depression in individuals who are avid social media users. The study compared the classification accuracy performance of Word2Vec paired with CNN and BERT based on data from social media users. Based on the accuracy analysis that was performed, the Word2Vec model with CNN had a precision of 52.1% with a loss of 69%, whereas the BERT model had a precision of 93% with a loss of 19%. The observations showed that the Transformer's encoder could perform the next sentence prediction randomly, allowing the model to maintain context representations of the tested content.

The above mentioned studies have demonstrated the use of deep learning models such as LSTM with attention mechanism and BERT-based models to detect depression in social media or text-based content. The Transformer's encoder was also observed to maintain context representations, allowing for improved next sentence prediction. Together, these studies demonstrate the potential for deep learning models in detecting mental health issues in social media content. However, based on a comprehensive review of the existing literature related to the application of deep learning and machine learning techniques in the context of mental health treatment, no prior studies have specifically examined the comparative effectiveness of LSTM classifications in comparison with the BERT classification model as well as the combination of LSTM models with Transformer word representation models, specifically the MentalBERT word representation model that have been designed to the mental health related domain.

Therefore, this research will conduct experiments to compare the already widely developed LSTM classification methods, BERT classification model and the combination of LSTM with the Transformer-based word representation model.

3. Materials and Methods

3.1. Data Collection

The data used in this study will be separated into training and testing data. The dataset was obtained from the CLPsych website (https://seanmacavaney.github.io/clpsych2021-shared-task/) that was collected around 2018. It consisted of tweets labeled as tweets with depressive disorder and non-depressive disorder. The total number of tweets that were collected in the dataset was 3200 tweets, consisting of 843 tweets with depressive disorder label and 2357 users with non-depressive disorder

labels, which can be represented in Table 1.

Table 1:Distribution of The Labelled Data

Tweets with Depressive Disorder	2357
Tweets with Non-Depressive Disorder	843

The labels used for categorizing users' tweets were 1 for Tweets with Depressive Disorder and 0 for Tweets with Non-Depressive Disorder, as shown in Table 2's sample data. A professional Psychologist was employed for analysing and validating the label appropriateness of the dataset. Over 344 data tweets were randomly chosen as samples in order to achieve a 95% confidence level and a 5% error rate of the analysis. Based on the analysis of the sampled data, the dataset can provide a sufficient representation of the behavior of people suffering from depressive disorder, allowing it to be used as initial data for the depressive disorder classification model.

Table 2:Samples of the Dataset

Tweet	Target
"With all of this unnecessary family drama, I feel like moving far away and starting over again. From one thing to another I just feel #depressed. Hope I get through this"	1
"Annoyed, sad, disappointed. Can anything go right lately.? #depressed #sad #done"	1
"Nobody chooses to be #depressed. But we do have the power to choose how we deal with it. We can give in & let the black dog hound us. Or we can choose to fight the inner b*tch that keeps us prisoner. Start by choosing to ignore her voice & choose #positive thoughts instead. pic.twitter.com/qQNAgbgC1I "	0
"I'm going to keep banging on about this, cos it's true. What you focus on, you get more of. Stop telling yourself you're #depressed or #anxious. Tell yourself you're happy, strong, confident, powerful. Not only cos you ARE, but cos your brilliant mind listens to what you tell it. pic.twitter.com/gBQn7yEjsJ"	0

3.2. Data Pre-Processing

The data pre-processing stage was carried out by processing text data into a numerical representation format that could be processed by a classification model.

3.2.1. Removal of URL Address

Before being processed by the word representation model, all URL addresses in the text data were removed. The URL text on a tweet does not provide information relevant to the context related to depressive disorder, so it could become noise in the text classification process.

3.2.2. Transformation to Lowercase

The process performed at this stage was to change the character at the beginning of each word of a sentence into a small letter. Converting text into lowercase format was done for the purpose of limiting the quantity of the vocabulary and minimizing the risk of noise and model overfitting.

3.2.3. Tokenization

Tokenization is the process of dividing a text into discrete words or tokens. Tokenization was employed to separate each word in a sentence, making it possible to spot patterns and relationships between words, which is essential for comprehending the context of depressive disorder and allowing for easier analysis and classification of the text.

Based on the embedding model utilized in this study, there were two different tokenization procedures. The Keras library was utilized for the Word2Vec embedding model. A total of 8536 unique words from the used dataset were obtained through the tokenization process.

For BERT and mBERT word embedding, textual data is processed using Bert Tokenizer, while MentalBERT word embedding uses AutoTokenizer for its tokenization process. Unlike Word2Vec, which processes text input data in word-by-word form, as stated by Devlin et al. (2018), the BERT model can process text input that is represented in the form of a sentence or a combination of sentences. These BERT tokenizer models will break down complete sentences into unique tokens; [CLS] at the start of each sentence to represent the sequence to be classified. [SEP] added between separated

sentences, or separator mark. [MASK] is randomly added in every sentence to be able to predict suitable words for those sentences.

3.2.4. Word Embedding

Word embedding is a term that represents a collection of language models and methods used for feature selection, commonly refered to as word representation (Li and Yang 2018). The main objective of this process is to convert text into low-dimensional vector form by mapping known words or phrases (Li and Yang 2018). In this work, two series of pre-processing stages were performed; pre-processing stages for the word embedding Word2Vec process, and pre-processing stages for the word embedding BERT, mBERT, and MentalBERT models.

3.2.4.1. Word2Vec

Word2Vec is a neural-network-based model with a technique that represents words into vector values, with the aim of obtaining words that are similar in use based on the closeness of the vector values, with a possibility of having more than one similarity value (multiple degrees of similarity) (Mikolov et al. 2013). The greater the similarity between the words, the closer the distance between the current vector values. By using the word's vector value, the word “king” can be found next to the vector word "man", or the word "woman", can be found around the vector “queen” value.

In the pre-processing stages for Word2Vec, the text data is first tokenized into individual words and then any punctuation marks are removed by using the class Tokenizer from the Keras library. The resulting list of words was then used to train the Word2Vec model, which assigned a vector value to each word based on its context within the text. This vector representation allows for efficient processing and comparison of large amounts of text data, making it a popular choice for natural language processing tasks.

3.2.4.2. BERT, mBERT, and MentalBERT

BERT (Bidirectional Encoder Representations from Transformers) is a language model developed by Google (Martínez-Castaño et al. 2021) which uses a deep neural network architecture to pre-train on large amounts of text data. The resulting model is able to generate contextual representations of words, meaning that it takes into account the surrounding words in a sentence when representing individual word. This allows for more accurate predictions and better performance on downstream natural language processing tasks.

The BERT component of the integrated model involves pre-trained BERT embeddings and fine-tuning (Devlin et al. 2018):

- BERT Pre-trained Embeddings
  BERT takes a sequence of words as input and produces contextual word embeddings for each word in the sequence.
- Fine-Tuning
  The fine-tuning process involves the acquisition of semantic meanings in order to enable models to effectively address specific problem domains. After obtaining BERT embeddings, the model fine-tunes these embeddings for the specific mental health classification task. Fine-tuning typically involves adding additional layers (e.g., fully connected layers) for classification on top of the BERT embeddings.

Multilingual BERT (mBERT) is an extension of BERT that has been trained on a multilingual corpus, making it useful for tasks that involve multiple languages. MentalBERT is a new variant of BERT that has been specifically designed to better understand human language and improve mental health care. It is expected from MentalBERT that it will be able to identify patterns and insights that can help clinicians in better understanding their patients' mental states and providing more effective treatment.

As BERT continues to evolve and improve, it is clear that it will play an increasingly important role in natural language processing in broader aspects. We implemented the "BERT-base-uncased" model for BERT, the "BERT-base-multilingual-uncased" model for mBERT, and the "mental-BERT-based-

uncased" model for the MentalBERT model.

Fig. 1 ilustrated the data pre-processing steps.

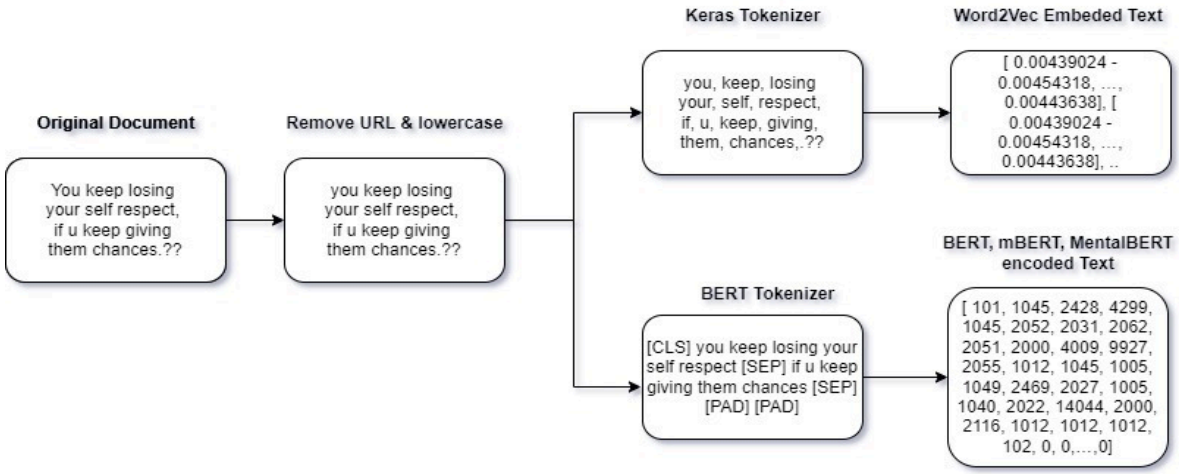

Fig. 1:Data Pre-Processing Ilustration

3.3. Depressive Disorder Classification Model

In this study, we sought to create classification models that could effectively determine whether a tweet qualifies as a depressive disorder or not. We trained 5 different classification models: LSTM initialized with pre-trained Word2Vec embeddings, BERT baseline classification model, LSTM with BERT pre-trained word representation, LSTM with mBERT, and LSTM with MentalBERT.

The LSTM unit in the integrated model can be represented as follows (Tadesse et al. 2020):

First gate of LSTM architecture is forget gate where input and previous hidden state are calculated using sigmoid function to decide which information is stored in memory cell.

$$ft = \sigma(Xt * Uf + Ht-1 * Wt) \quad (1)$$

$$Ct-1 * ft = 0 \ldots if\ ft = 0\ (Forget\ everything) \quad (2)$$

$$Ct-1 * ft = Ct-1 \ldots if\ ft = 1\ (Forget\ nothing) \quad (3)$$

$$it = \sigma(Xt * Ui + Ht-1 * Wi) \quad (4)$$

Next step is to calculate the importance of previous information using tanh layer.

$$Nt = tanh(Xt * Uc + Ht-1 * Wc) \quad (5)$$

Those 2 prior data then calculated as memory cell of new cell state, and then added to previous memory cell state ct-1 to update current cell state as ct.

$$Ct = (ft * Ct-1 + it * Nt)\ (updating\ cell\ state) \quad (6)$$

Then the output is obtained by calculating weight, input, and hidden state of previous processing while also producing new hidden state of current process to be used by next process of LSTM.

$$Ot = \sigma(Xt * Uo + Ht-1 * Wo) \quad (7)$$

$$Ht = (Ot * tanh(Ct)) \quad (8)$$

Fig. 2 and Fig. 3 presented the LSTM with Word2Vec and LSTM with BERT-based Transformer word representation model architecture respectively. As shown in Fig. 2, to handle the word vector input that has previously been processed using Word2Vec, an Embedding layer was added to the input

layer of the classification model. A layer of TFBertModel was added on the architecture of BERT classification model in Fig. 4, as well as the LSTM combination model combined with BERT-based Transformer model, in order to process the encoded BERT tokens.

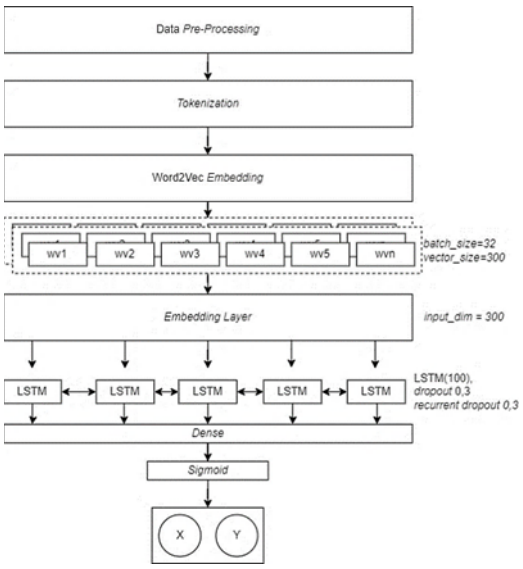

Fig. 2:LSTM with Word2Vec Architecture for Depressive Disorder Classification

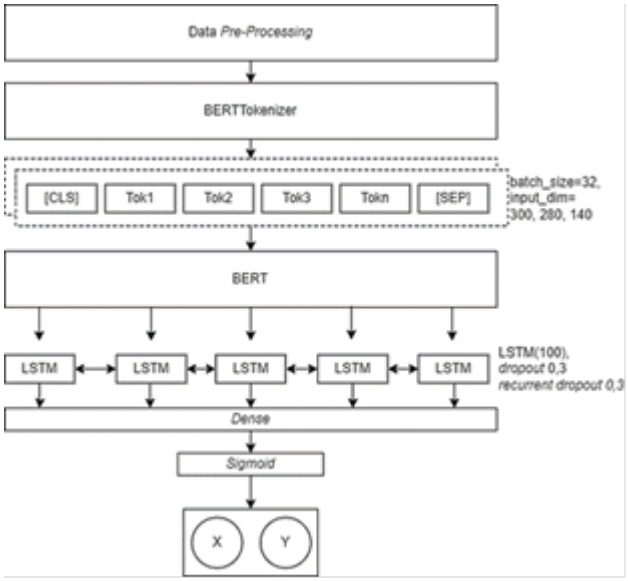

Fig. 3:LSTM with BERT-based Transformer Architecture for Depressive Disorder Classification

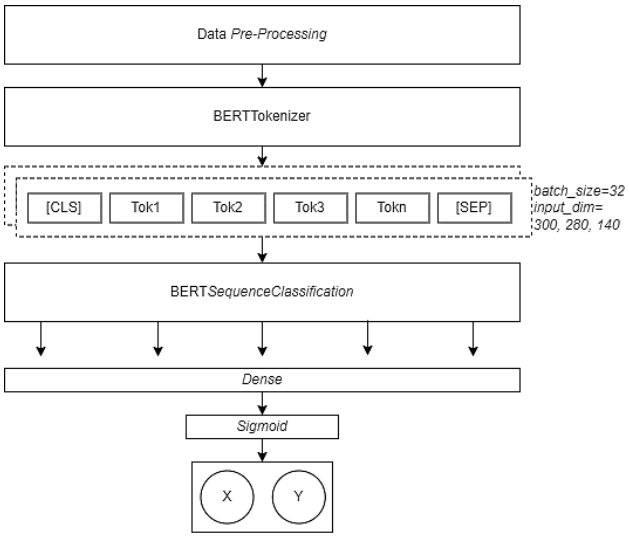

Fig. 4:BERT-base Classification Model Architecture for Depressive Disorder Classification

For the combined model LSTM, a single drop-out of 0.3 and a recurrent drop-out of 0.3 were implemented as additional parameters. In addition to layer drop-out, an early stop callback function was implemented to decrease the possibility of overtraining or overfitting on the model. This function will halt the training process when the training model has obtained an accuracy >0.95. In the case of binary classification, the Dense layer was added with an activation sigmoid.

After building the model architecture, the model was trained using the training dataset. Hyperparameters used for tuning on each model in this experiment were listed in Table 3. The training process was iterative, with the model being retrained and tuned multiple times until it achieved a satisfactory level of performance. Each model's training phase employed the Adam optimizer and a

separate validation dataset comprised of 0.2% of the training data, split randomly, to make sure it could accurately classify depressive disorders.

Table 3:List of Optimized Hyperparameters

Hyperparameter	**Values**
Number of epochs	[15, 20, 30]
Input dimention	[140, 280, 300]
Computing Units	[GPU, TPU]

Each model has a unique set of hyperparameters. The number of epochs in BERT and LSTM with BERT, MentalBERT, and mBERT were all less than 10 epochs as the training accuracy have reached above 0.95.

Table 4:The best configuration of hyperparameters for each models

Method	***input_dim***	***epochs***	***loss***	***val_loss***	***accuracy***	***val_accuracy***
BERT	300	5	0.0766	0.3373	0.9819	0.8242
LSTM - Word2Vec	300	30	0.1523	0.4165	0.9399	0.7695
LSTM - BERT	300	9	0.0880	0.4005	0.9673	0.8457
LSTM – MentalBERT	140	6	0.1041	0.4052	0.9653	0.8203
LSTM - mBERT	140	6	0.1294	0.3467	0.9585	0.8652

3.4. Evaluation

In this study, the accuracy, precision, recall, and F1-scores are compared to assess each model's performance. Accuracy is calculated by dividing the accurate prediction value with the total number of the dataset (Emmanuel, Folorunso, and Thomas 2020), as in (9). Precision is define as the amount of labeled data identified as true positives (Emmanuel, Folorunso, and Thomas 2020), which is the ratio of positive samples predicted accurately to the number of examples classified as positive labels, presented in (10). Recall is defined as the proportion of true positives that has been classified with the total number of examples labeled as positive (Karthika, Murugeswari, and Manoranjithem 2019), and F1-Score is the arithmetic mean of precision and recall (Karthika, Murugeswari, and Manoranjithem 2019). The higher the F1 score, the better the classification performance (Emmanuel, Folorunso, and Thomas 2020), as shown in (11) and (12).

$$Accuracy = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \tag{9}$$

$$Precision = \frac{TP}{TP+FP} \tag{10}$$

$$Recall = \frac{TP}{TP+FN} \tag{11}$$

$$F1 - Score = 2\,x\left(\frac{Precision\ x\ Recall}{Precision+Recall}\right) \tag{12}$$

Where:

TP = True Positive

TN = True Negative

FP = False Positive

FN = False Negative

The models with the highest ratings are shown in bold.

Table 5:Model Performance

METHOD	*ACCURACY*	*PRECISION*	*RECALL*	F1-*SCORE*	TRAINING TIME
BERT	0.82	0.76	**0.80**	0.78	40 – 90 SECONDS
LSTM - WORD2VEC	0,75	0,70	0,72	0,71	90 SECONDS
LSTM - BERT	0,84	0,78	0,77	0,77	40 – 90 SECONDS
LSTM – MENTALBERT	**0,86**	**0,81**	0,76	**0,78**	20 - 60 SECONDS
LSTM - MBERT	0,81	0,74	0,72	0,73	70 - 115 SECONDS

Below are the confusion matrices visualizing the model predictions.

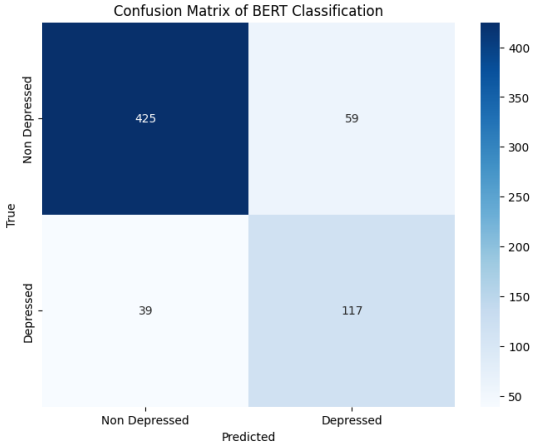

Fig. 5: Confusion Matrix of BERT Base Classification

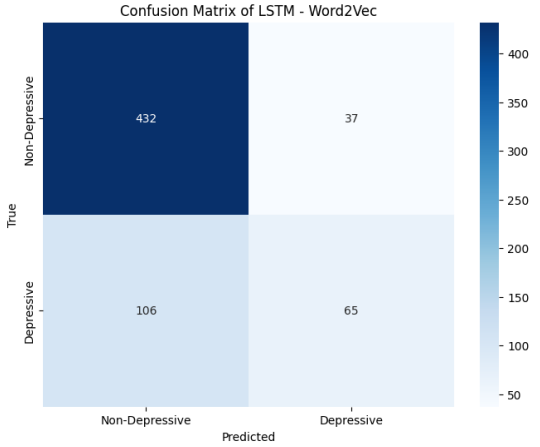

Fig. 6: Confusion Matrix of LSTM-Word2Vec Classification

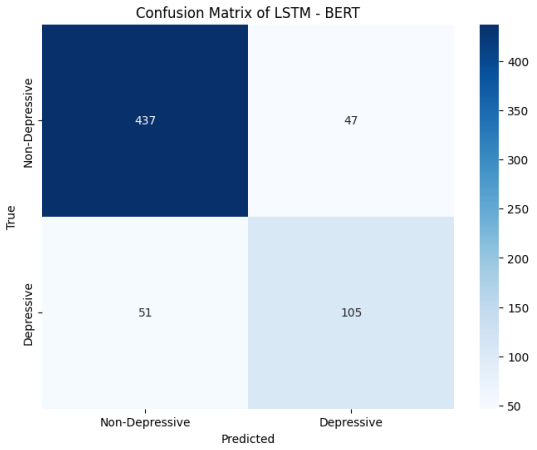

Fig. 7: Confusion Matrix of LSTM-BERT Classification

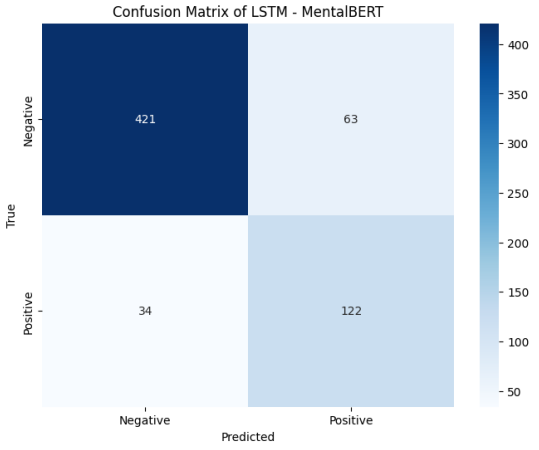

Fig. 8: Confusion Matrix of LSTM-MentalBERT Classification

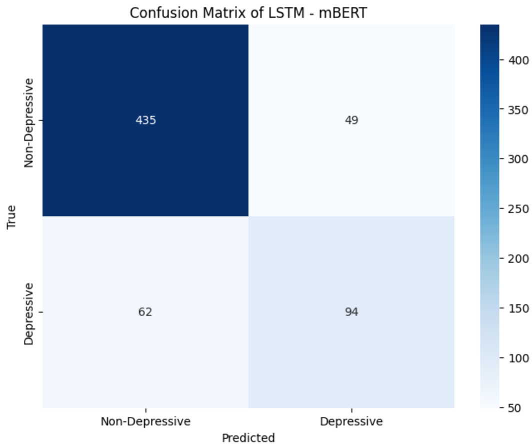

Fig. 9: Confusion Matrix of LSTM-mBERT Classification

Models that used the transformer's word representation technique performed slightly better than the ones that did not. Specifically, the LSTM with Transformer-based MentalBERT model achieved the best accuracy of 86%, precision of 81%, recall of 76%, and F1-score of 78%. We can also see from the illustration of the confusion matrix above that the combined LSTM with MentalBERT can detect true depressive disorder more than the other models. On the other hand, the non-transformer models, in this case, the LSTM with Word2Vec, had an accuracy of 75%, precision of 70%, recall of 72%, and F1 -score of 71%.

Here are some examples of the prediction output from LSTM-MentalBERT model as presented in Table 6.

Tweet	**Target**	**Prediction**
Example of Depressive Data index: 361 Text: [CLS] smiling had been always natural for me. it was easy to conjure up on my face. now it becam Encoded Text: tf.Tensor([101 5629 2018 2042 2467 3019 2005 2033 1012 2009 2001 3733 2000 9530 25243 2039 2006 2026 2227 1012 2085 2009 2150 2028 1997 1996 18263 2477 2000 2079 1999 2026 2166 1010 2279 2000 2542 1012 102 0], shape=(300,), dtype=int64) True Label: 1 Predicted Label: [1.]	1	1

```
Example of Non-Depressive Data
index: 93
Text: [CLS] chris is leaving and dad isn't here and made isn't here to clean and the weather isn't super amazin
Encoded Text: tf.Tensor(
[ 101 3782 2003 2975 1998 3611 3475 1005 1056 2182 1998 2081 3475 1005
 1056 2182 2000 4550 1998 1996 4633 3475 1005 1056 3565 6429  102    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0], shape=(300,), dtype=int64)
True Label: 0
Predicted Label: [0.]
```

0	0

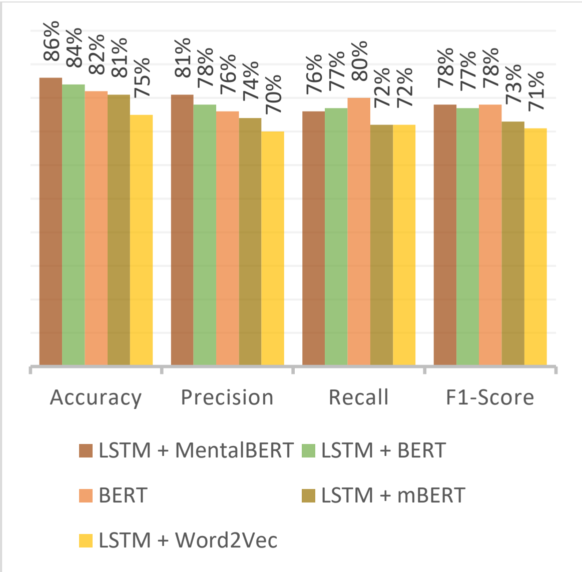

Fig. 10: Performance Comparison

4. Result and Discussion

Similar to the study conducted by Winata et al. (2018), the proposed model of LSTM with Word2Vec also experiences problems in the context of word structures with more complicated grammar. This might be a result of Word2Vec's corpus having a different word structure than the expression that was mostly used on Twitter. With a vocabulary of up to 100 billion words and about 3 million unique terms, Word2Vec is establishing a distinctions model based on data from Google News (Winata, Kampman, and Fung 2018). However, as some of the grammar used by those who expressed their emotions on Twitter occasionally deviated from that of conventional terms, there are variances in the new word's form and structure.

When compared to the BERT baseline classification model, the LSTM model paired with the Transformer models performed slightly better. BERT model predicts the next word based on the context of the phrase (Devlin et al. 2018), whereas the LSTM model handles sequential data by employing memory units to recall past information and decide how much of that information will be sent to the next sequence (Greff et al. 2017). This demonstrates that the LSTM model outperforms the BERT model in terms of representing sequences or data paths in text.

We also conducted a comparative analysis between our proposed system and existing state-of-the-art systems that were specifically developed for the purpose of analyzing and predicting mental health issues based on data from social media platforms. The results of this comparison are presented in Table 7.

Table 7: comparison with prior methods for mental health classification on social media

Author	Social Media	Classification Methods	Accuracy
Tadesse et al. (2020)	REDDIT	LSTM-CNN with Word2Vec	93.8%
Kim et al., (2020)	Reddit	XGBoost-TF-IDF	Depression 71.69% Anxiety 70.41% Bipolarity 85.53% BPD 85.14% Schizophrenia 86.72% Autism 94.91%
		CNN-Word2Vec	Depression 75.13% Anxiety 77.81% Bipolarity 90.20% BPD 90.49% Schizophrenia 94.33% Autism 96.96%
Syed and Narvekar (2021)	Twitter	BERT	93%
		CNN-Word2Vec	52.1%
Ghosh et al. (2021)	Twitter	LSTM-CNN	99.42%
Almars (2022)	Arabic Social Media Content	Bi-LSTM	83%
Our Conducted Methods	Twitter	BERT	82%
		LSTM - Word2Vec	75%
		LSTM - BERT	84%
		LSTM - MentalBert	86%
		LSTM - mBERT	81%

Tadesse et al. (2020) applied an LSTM-CNN model with Word2Vec embeddings on Reddit data, achieving an accuracy of 93.8%. Kim et al. (2020) employed XGBoost-TF-IDF and CNN-Word2Vec for classification on Reddit, showcasing decent performance in distinguishing depression (71.69%), anxiety (70.41%), bipolarity (85.53%), borderline personality disorder (BPD) (85.14%), schizophrenia (86.72%), and autism (94.91%).

In comparison, Syed and Narvekar (2021) utilized BERT for Twitter data classification, reaching

an accuracy of 93%. Ghosh et al. (2021) achieved remarkable accuracy of 99.42% using an LSTM-CNN model on Twitter data. Almars (2022) explored Arabic social media content and reported an accuracy of 83% using a BI-LSTM model.

Our conducted methods on Twitter data produced competitive results, with BERT achieving 82% accuracy. The LSTM with Word2Vec embeddings reached 75%, while the LSTM with BERT achieved an accuracy of 84%. Notably, the LSTM with MentalBERT outperformed others with an accuracy of 86%, demonstrating its potential in mental health classification. Finally, the LSTM with mBERT achieved an accuracy of 81%. These findings highlight the effectiveness of different classification methods, with our approach showcasing promising results in the challenging task of mental health classification on social media platforms.

The LSTM-MentalBERT architecture excels due to its ability to leverage the complementary strengths of LSTM and BERT. MentalBERT, a variant of BERT fine-tuned for mental health-related text, is specifically trained to the domain, enhancing its performance on mental health classification tasks (Ji et al. 2022). LSTM models can maintain the relationship between words for a longer period using memory cells, as they have memory cells inside the architecture to keep prior processing output and then use it as extra input consideration for the current process and keep updating it in every process that occurs, making the architecture very good for sequential data, for example, sentences in tweets (Greff et al. 2017). LSTM, with its proficiency in capturing sequential patterns and local context, complements MentalBERT's domain-specific pre-training and fine-tuned understanding of mental health language, as It has been trained and allocated to the body of data related to mental health problems, including anxiety disorder, depressive disorders, stress, bipolar, to PTSD (Ji et al. 2022). While LSTM excels at recognizing nuanced emotional expressions within sentences, MentalBERT's domain-specificity ensures a profound grasp of mental health terminology and context. This combination empowers the model to excel in the nuanced domain of mental health classification, where context, sensitivity, and interpretability are paramount. The balance between global and local context processing makes this architecture a robust choice for understanding and supporting mental well-being through AI-driven text analysis.

Our research has demonstrated the promising performance of the classification model in the context of mental health analysis, but there are still certain limitations to keep in mind. Firstly, the performance of the classification model is significantly influenced by the quality and representativeness of the training dataset. Therefore, the importance of having a robust and diverse dataset for model initialization cannot be overstated. Ensuring that the training dataset accurately reflects the population under study is pivotal in achieving meaningful results. Moreover, the training process itself demands continuous updates with newer and more relevant data to maintain model relevance, which could pose challenges in areas such as sampling, optimization of prediction approaches and their characteristics, generalizability, privacy, and other ethical concerns.

A notable limitation of our study pertains to the significant computational resources and time required for model training to achieve the desired accuracy threshold of >95%. We discovered considerable variances in the average training time required per epoch depending on the word embedding we used. Our research found that models using MentalBERT embeddings performed well in training, frequently taking fewer than ten epochs to obtain an accuracy of more than 95%. This resulted in training durations as short as 20 to 60 seconds per epoch for MentalBERT, and an average of 40 to 90 seconds for models utilizing BERT embeddings. However, while mBERT gave great contextual comprehension, it required substantially longer training sessions, exceeding 115 seconds per epoch, while reaching the target accuracy level after 8 epochs.

In comparison, the Word2Vec embedding method provided a more significant barrier in terms of training time, needing an average of 90 seconds per epoch. Despite these longer training intervals, Word2Vec-based models did not manage to achieve the critical accuracy threshold of greater than 95%,

even after 30 epochs of training.

Furthermore, Word2Vec embeddings were unable to use GPU acceleration and had to rely completely on CPU processing, which contributed to their lengthy training periods. BERT, mBERT, and MentalBERT embeddings, on the other hand, were tuned for GPU utilization, notably on GPU V100, resulting in more efficient training but also requiring more significant computing resources.

Results obtained from this research suggest that incorporating transformer-based word representations with LSTM can improve the performance of classification models for mental healthcare domain-specific tasks. It utilizes the strengths of both transformer and LSTM models to interpret the syntactic and contextual information of each word and represent an effective sequence of sequential data. Overall, our findings highlighted the importance of selecting appropriate word representation techniques for deep learning classification models in natural language processing tasks. These findings have important implications for the development of more effective tools for mental health professionals to analyze and understand their patients' language, as well as for researchers studying the intersection of language and mental health. As the field continues to evolve, it will be important to continue exploring new approaches to improve the accuracy and applicability of these models in real-world settings. Ultimately, this research has the potential to contribute to better understanding and treatment of mental health conditions through a more accurate analysis of patient language.

5. Conclusion

This research has presented methods for classifying tweets with depressive disorder over the collected dataset from the Twitter social network. As conclusion, our proposed architecture performed best when combining an LSTM classification model with a Transformer MentalBERT word representation, outperforming the other models achieved a notable 86% increase in terms of accuracy.

This improvement in accuracy is particularly significant in light of the challenges posed by the complexities of word structures, especially in the context of social media language, as observed in a study by Winata et al. (2018). We attribute some of these challenges to the variations between Word2Vec's corpus and the expressions predominantly used on Twitter, leading to differences in word structures. Our findings underscore the critical role of selecting appropriate word representation techniques for deep learning models in natural language processing tasks.

The experiment results and the evaluation test of the classification model showed that the LSTM classification model achieved a significant boost in accuracy when combined with Transformer word representation models. It utilizes the strengths of both transformer and LSTM models to interpret the syntactic and contextual information of each word and represent an effective sequence of sequential data. However, the accuracy improvement may not solely be attributed to the combination of transformer and LSTM models, as other factors such as dataset size, preprocessing techniques, computational resources, and hyperparameter tuning could have also contributed to the observed results. Additionally, investigating the influence of different word representations and classification algorithms would further contribute to a more thorough analysis of overall model performance in real-world applications. By exploring different variations of models and datasets, we can gain a comprehensive understanding of their impact on performance.

References

Almars, Abdulqader M. 2022. "Attention-Based Bi-LSTM Model for Arabic Depression Classification." *Computers, Materials and Continua* 71(2): 3091–3106.

Alsentzer, Emily et al. 2019. "Publicly Available Clinical BERT Embeddings." http://arxiv.org/abs/1904.03323.

Asad, Nafiz Al, Md Appel Mahmud Pranto, Sadia Afreen, and Md Maynul Islam. 2019. "Depression

Detection by Analyzing Social Media Posts of User." *2019 IEEE International Conference on Signal Processing, Information, Communication and Systems, SPICSCON 2019*: 13–17.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *Naacl-Hlt 2019* (Mlm): 4171–86.

Emmanuel, Femi, Olusegun Folorunso, and Friday Thomas. 2020. "Machine Learning Techniques for Hate Speech Classification of Twitter Data : State-of-the-Art , Future Challenges and Research Directions." *Computer Science Review* 38: 100311. https://doi.org/10.1016/j.cosrev.2020.100311.

Ethayarajh, Kawin. 2019. "How Contextual Are Contextualized Word Representations? Comparing the Geometry of BERT, ELMO, and GPT-2 Embeddings." *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*: 55–65.

Ghosh, Tapotosh, Hasan Al Banna, Jaber Al Nahian, and Kazi Abu Taher. 2021. "A Hybrid Deep Learning Model to Predict the Impact of COVID-19 on Mental Health Form Social Media Big Data." (June): 1–21.

Greff, Klaus et al. 2017. "LSTM: A Search Space Odyssey." *IEEE Transactions on Neural Networks and Learning Systems* 28(10): 2222–32. http://ieeexplore.ieee.org/document/7508408/.

Ji, Shaoxiong et al. 2022. "MentalBERT : Publicly Available Pre-Trained Language Models for Mental Healthcare." (June): 7184–90.

Karthika, P., R. Murugeswari, and R. Manoranjithem. 2019. "Sentiment Analysis of Social Media Network Using Random Forest Algorithm." *IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing, INCOS 2019*: 1–5.

Kim, Jina, Jieon Lee, Eunil Park, and Jinyoung Han. 2020. "A Deep Learning Model for Detecting Mental Illness from User Content on Social Media." 10: 11846. https://doi.org/10.1038/s41598-020-68764-y.

Lee, Jinhyuk et al. 2020. "BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining." *Bioinformatics* 36(4): 1234–40.

Li, Yang, and Tao Yang. 2018. "Guide to Big Data Applications." (June).

Lin, Lin, Xuri Chen, Ying Shen, and Lin Zhang. 2020. "Towards Automatic Depression Detection: A BiLSTM/1D CNN-Based Model." : 1–20.

Martínez-Castaño, Rodrigo, Amal Htait, Leif Azzopardi, and Yashar Moshfeghi. 2021. "BERT-Based Transformers for Early Detection of Mental Health Illnesses." In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, , 189–200. https://link.springer.com/10.1007/978-3-030-85251-1_15.

Miaschi, Alessio, and Felice Dell'Orletta. 2020. "Contextual and Non-Contextual Word Embeddings: An in-Depth Linguistic Investigation." *Proceedings of the Annual Meeting of the Association for Computational Linguistics*: 110–19.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*: 1–12.

Minaee, Shervin et al. 2020. "Deep Learning Based Text Classification: A Comprehensive Review." 1(1): 1–43. http://arxiv.org/abs/2004.03705.

Pavlova, Alina, and Pauwke Berkers. 2020. "Mental Health Discourse and Social Media: Which

Mechanisms of Cultural Power Drive Discourse on Twitter." *Social Science and Medicine* 263: 113250.

Reynard, Darcy, and Manish Shirgaokar. 2019. "Harnessing the Power of Machine Learning: Can Twitter Data Be Useful in Guiding Resource Allocation Decisions during a Natural Disaster?" *Transportation Research Part D: Transport and Environment* 77: 449–63.

Santomauro, Damian F. et al. 2021. "Global Prevalence and Burden of Depressive and Anxiety Disorders in 204 Countries and Territories in 2020 Due to the COVID-19 Pandemic." *The Lancet* 398(10312): 1700–1712.

Sarsam, Samer Muthana et al. 2021. "A Lexicon-Based Approach to Detecting Suicide-Related Messages on Twitter." *Biomedical Signal Processing and Control* 65(March): 102355. https://doi.org/10.1016/j.bspc.2020.102355.

Syed, Maharukh, and Meera Narvekar. 2021. "A Decision Support System for Predicting Socially Depressed Users Using Bidirectional Encoders Representations from Transformers (BERT)." 23(3): 209–13.

Tadesse, Michael Mesfin, Hongfei Lin, Bo Xu, and Liang Yang. 2020. "Detection of Suicide Ideation in Social Media Forums Using Deep Learning." *Algorithms*: 1–19.

WHO. 2020. "Depression." https://www.who.int/news-room/fact-sheets/detail/depression (February 2, 2021).

Winata, Genta Indra, Onno P. Kampman, and Pascale Fung. 2018. "Attention-Based LSTM for Psychological Stress Detection from Spoken Language Using Distant Supervision." *IEEE International Conference on Acoustics, Speech and Signal Processing*.

Zeberga, Kamil et al. 2022. "A Novel Text Mining Approach for Mental Health Prediction Using Bi-LSTM and BERT Model." *Computational Intelligence and Neuroscience* 2022.

Zhou, Tie Hua, Gong Liang Hu, and Ling Wang. 2019. "Psychological Disorder Identifying Method Based on Emotion Perception over Social Networks." *International Journal of Environmental Research and Public Health* 16(6).