Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Depression Detection on Mandarin Text through Bert Model

Yung Teck Kiong[1], Cheah Wai Shiang[1,*], Mahir Pradana[2], Hamizan Sharbiniung[1], Iwan Tri Riyadi Yanto[3]

[1] Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 93250 Kota Samarahan, Sarawak, Malaysia
[2] Department of Business Administration, Telkom University, Jalan Terusan Buah Batu 1, Bandung 40257, Indonesia
[3] Departement of information system, Fakulty of technology and Applied Science, Universitas Ahmad Dahlan, Kota Yogyakarta, Daerah Istimewa Yogyakarta 55166, Indonesia

ABSTRACT

Depression is currently one of the most prevalent mental disorders and its incidence has been rising significantly in Malaysia amid the Covid-19 pandemic. While previous studies have demonstrated the potential of artificial intelligence technology in analysing social media texts to detect signs of depression, most of these studies have focused on English textual content. Considering that Mandarin is the second most widely spoken language worldwide, it is worthwhile to explore depression detection techniques specifically tailored for Mandarin textual content. This research aims to examine the effectiveness of the BERT model in text classification, particularly for detecting depression in Mandarin. The study proposes the utilization of the BERT model to analyse social media posts related to depression. The model is trained using the WU3D dataset, which comprises a collection of over 2 million text data sourced from Sina Weibo, a prominent Chinese social media platform. Given the dataset's inherent imbalance, text augmentation techniques were employed to assess whether they contribute to improved model performance. The findings suggest that the BERT model trained on the original dataset outperformed the model trained on the augmented dataset. This implies that the BERT model is well-equipped to handle imbalanced datasets effectively. Furthermore, it is speculated that the augmented dataset did not introduce novel information or knowledge during the model training process. Notably, the highest-performing model achieved an impressive accuracy rate of 88% on the testing dataset.

Keywords:
NLP; Depression; Machine learning; Transformer; BERT

1. Introduction

Depression has emerged as a prevalent global ailment, affecting a substantial number of individuals. The World Health Organization (WHO) reports that approximately 280 million people, accounting for 3.8% of the global population, are currently grappling with depression [1]. Depression symptoms include feeling down, losing interest in activities, difficulty concentrating, feeling helpless

[*] *Corresponding author.*
E-mail address: c.waishiang@gmail.com

https://doi.org/10.37934/araset.60.2.295311

or hopeless and even having thoughts of suicide. Depression can be caused by various factors, such as overall health, quality of life [2], family history, genetics [3] and more.

In Malaysia, the prevalence of depression has been on the rise, particularly during the COVID-19 pandemic and the subsequent lockdown period. A study has revealed that during the early COVID-19 lockdown, approximately 11.1% of healthcare workers reported experiencing suicidal thoughts [4]. In addition, the prevalence of depression with comorbid anxiety was 19.4% among Chinese adults in Malaysia [5].

In today's information-driven era, social media has become a platform that enables people to connect and communicate with each other regardless of time and space. It allows individuals to effortlessly access information and share their thoughts, feelings and ideas. Social media serves as a platform for expressing emotions and thoughts, seeking support and fostering connections, transcending geographical boundaries [6]. When used appropriately, social media can have a positive impact on its users. However, it is often blamed as one of the contributing factors to depression, particularly among teenagers and young adults. Research suggests that the amount of time spent on social media, specific activities, emotional investment and addiction to social media are strongly associated with depression [7].

To detect depression, various techniques have been introduced, including linear regression and different learning models like recurrent neural networks and convolutional neural networks. Many researchers have conducted sentiment analysis on English textual content [8-10]. However, research focusing on sentiment analysis of Mandarin text is relatively limited [11]. Given that Mandarin is the second most widely spoken language globally, it is crucial to explore and develop early depression detection techniques for Mandarin text, particularly for the benefit of Mandarin speakers.

1.1 Technique to Detect Depression on Social Media

Social media platforms have become a prominent means for individuals to express their thoughts, ideas, feelings and opinions through text, images and videos. As a result, social media has emerged as a rich source of data for studying and analysing user behaviour. Several studies have been conducted to detect or analyse depression using text from social media.

One study focused on Facebook comments to explore and detect depressive behaviour in users [12]. The researchers employed supervised learning approaches, such as decision trees, k-nearest neighbours, support vector machines (SVM) and ensemble classifiers, for the task of depression detection. They collected a total of 7,145 English comments from Facebook, with 4,149 comments indicated as depressive and 2,996 comments as non-depressive. The study found that 54.77% of depressive comments were captured between midnight and midday, suggesting a possible influence of time on depressive emotions. The accuracy of the classifiers ranged between 60% and 80%.

Another study focused on Twitter data and compared various classification algorithms for depression detection [13]. The study examined linear regression, SVM, multilayer perceptron, decision trees, random forests, adaptive boosting, bagging predictors and gradient boosting. The authors also investigated the impact of sample balancing on the dataset and found that certain classifiers benefited from balanced sampling, while others did not. This highlights the importance of experimenting with data balancing techniques to improve classification accuracy.

Deep learning approaches have also been utilized for depression detection. One study compared Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) models [14]. While RNN is well-suited for sequential data like text sentences and audio, CNN showed superior performance in natural language processing (NLP) tasks.

A study conducted on the Mandarin language using Sina Weibo, a popular social media platform in China, employed a multimodal learning-based deep neural network model called Multimodal Feature Fusion Network (MFFN) for depression detection [15]. The authors contributed a large-scale labelled dataset, the Weibo User Depression Detection Dataset (WU3D), containing over 400,000 depression-labelled data and 1,000,000 non-depression data.

In another study, CNN and RNN methods were applied to detect depression tendencies in Chinese text [16]. The proposed model utilized parts of the CNN structure, including embedding, convolutional layers, pooling layers and max-pooling layers, for feature extraction. The subsequent layer involved LSTM for feature data processing, dropout and an output layer for result production. The study showed that the proposed model outperformed an ordinary CNN model by 5% and highlighted the potential for further experimentation with hyperparameters.

Furthermore, image processing techniques have been employed for depression prediction on Instagram [17]. Unlike text-based platforms, Instagram is primarily focused on photo and video sharing. In this study, an application was designed to collect photos and history from participants' Instagram accounts. The collected data were then used to train a 100-tree random forest classifier algorithm.

Another study proposed a multimodal dictionary learning solution for depression detection by continuously harvesting social media content [18]. The authors constructed three sets of datasets, including a Depression Dataset (D1) of tweets from depressed users, a Non-depressed Dataset (D2) from non-depressed users and a Depression-candidate Dataset (D3) containing loosely related tweets. Various features such as social network features, user profile features, visual features, emotional features, topic-level features and domain-specific features were extracted and used for classification.

In conclusion, depression detection can be approached through various methods, including NLP, image processing and multimodal analysis. While many studies have focused on NLP-based approaches, there is a need to explore the capabilities of NLP models for depression detection in Mandarin and investigate attention mechanism-based models like BERT for this task.

1.2 Background Studies on Algorithms used for Depression Detection

Natural language, including languages like English, Mandarin and others spoken by humans, serves as a means for human communication. The field of natural language processing (NLP) aims to equip computers with the ability to comprehend texts and speech in a manner similar to humans. However, NLP poses significant challenges. In contrast to other machine learning tasks such as computer vision and classification, NLP models tend to be large in size. This is primarily because natural language data is vast, often approaching infinite quantities. Additionally, many words in natural language exhibit ambiguity, possessing multiple meanings. Furthermore, natural language exhibits feature such as homophones, sarcasm, metaphors, grammar and homonyms. Consequently, developing a machine capable of comprehending human language on par with humans is an immensely difficult task [22].

The recurrent neural network (RNN), introduced in 1986, has become a highly popular model in the field of natural language processing (NLP) [23]. It is an evolution of feedforward neural networks, including linear regression and convolutional neural networks (CNN). RNN stands out by utilizing internal memory, allowing it to retain information about the context of input data and make predictions about what might follow. This capability is particularly valuable when dealing with sequential data, as considering the context is vital for generating meaningful and relevant predictions. In many NLP tasks, the RNN model frequently outperforms other models, such as CNN

[24]. Nonetheless, one notable challenge with RNNs is the issue of gradient vanishing or exploding during the backpropagation process, which can impede their effectiveness.

In 1997, Sepp Hochreiter and Jürgen Schmidhuber introduced a significant advancement to the concept of recurrent neural networks (RNN) with the development of long short-term memory (LSTM) [25]. LSTM builds upon the foundation of RNNs by incorporating additional components known as gates within each layer. These gates include the input gate, forget gate and output gate. The input gate determines which values from the input or previous layer should be used to modify the memory. The forget gate, as the name implies, decides which information should be discarded. Lastly, the output gate allows the processed data within a layer to be outputted to the next layer.

One notable advantage of LSTM is its ability to address the issue of gradient vanishing encountered in traditional RNNs. By incorporating these gates, LSTM provides better memory performance and is particularly suited for processing sequential inputs of variable lengths. This feature allows LSTM to effectively capture and retain essential information from the input data, regardless of the sequence's length.

Seq2Seq, also known as the encoder-decoder model, is a machine learning approach widely used in NLP tasks. It takes an input sequence and produces an output sequence. It finds applications in machine translation, image captioning, text summarization and more [26]. The encoder, typically built with a stack of RNN models like LSTM, encodes the input sequence. The decoder, also constructed with RNN models, decodes and generates the output sequence. The encoder vector, a hidden state, acts as the initial state for the decoder, encapsulating the encoder's information and facilitating the generation of the output sequence.

The attention mechanism, introduced in 2014 by Dzmitry Bahdanau *et al.,* [27], was developed to overcome the limitation of using a fixed-length vector in encoder-decoder models, where information could potentially be lost between the encoder and decoder. In traditional encoder-decoder models, the decoder relies on a fixed-length vector generated by the encoder to make predictions. However, this approach becomes problematic when dealing with long input sequences, as crucial information can be lost.

The attention model addresses this issue by predicting each output word based on a specific part of the input sequence that is considered the most relevant, rather than relying on the entire context. By focusing on the most relevant information at each step, the attention mechanism ensures that important details are not overlooked or omitted during the decoding process. This approach enhances the model's ability to capture and utilize all relevant information, leading to improved performance and more accurate predictions.

The Transformer model, introduced by Ashish Vaswani *et al.,* [28], relies solely on the attention mechanism. It eliminates the need for recurrent or convolutional neural networks and instead utilizes self-attention to capture word relationships within a sequence. This approach enables the Transformer to achieve superior performance compared to previous models in machine translation. Additionally, the Transformer has found successful applications in computer vision tasks such as object classification, resolution enhancement and video processing [29-32].

BERT (Bidirectional Encoder Representations from Transformers) is a pre-training Transformer-based model that learns contextual relationships in sequential data from both directions. Unlike the original Transformer, BERT focuses on generating a generalized model instead of accomplishing specific tasks. Pre-trained BERT models are readily applicable to various tasks with minimal fine-tuning, saving computation resources and time. BERT employs the Masked Language Model (MLM), which allows it to learn the contextual relationships in sequential data from both directions. This approach mimics human reading behaviour, as reading a sentence from both directions can help algorithms better understand the context and make more accurate sense of the text.

These studies highlight the effectiveness of machine learning in analysing and detecting depression using various methods. NLP is commonly used for textual-based depression detection, although most studies focus on English. Exploring the attention mechanism, particularly the BERT model, could enhance accuracy in this area. Further research in this direction holds promising potential for advancements in depression analysis and detection.

This research aims to assess the performance of the BERT model in text classification, specifically in Mandarin. The study proposes the use of the BERT model for detecting depression on social media. The model is trained using the WU3D dataset, which comprises over 2 million text data collected from a popular Chinese social media platform called Sina Weibo. Considering the dataset's imbalance, text augmentation techniques were employed to examine whether they can enhance the model's performance.

2. Methodology

2.1 Dataset Preparation and Processing

Dataset WU3D published by Wang *et al.,* [15] will be used in the study. WU3D includes two JSON files that are for depressed data and non-depressed data and contains various attributes such as username, gender, birthday, post content and other post-related attributes like several sharing and like. There are more than 400000 depressing data and 1700000 non-depressed data that are collected from Sina Weibo since the Year 2020.

The data is stored in JSON format and grouped by user. In other words, a user could contain one or multiple post data. However, this study is not concerned with analysing user behaviour and individual depression tendency. Thus, user information and non-content post attributes will be excluded. Finally, “tweet_content” is the only attribute that will be retained.

The dataset is considered clean as all of the non-text contents had been removed by the authors, thus there is no further cleaning processing for this dataset. But, since only the attribute “tweet_content” is needed, the data need to be transformed into table format and stored in CSV format for ease of model training process. The depressed and non-depressed datasets will be transformed and combined into a single dataset with an additional attribute called “depressed” (1 for depressed, 0 for non-depressed) to indicate whether it is depressed or not.

However, there is another issue that needs to be addressed in this dataset. This dataset is unbalanced. Non-depressed data is significantly more than depressing data. Thus, a comparison of sampling techniques needs to be done to determine a better solution that could improve the detection accuracy.

2.2 Establishing Model

Before BERT was introduced, the model is trained to solve specific NLP tasks such as sentiment analysis, text classification, document summarization and etcetera. Each trained model is not capable of solving other tasks.

As mentioned in the previous chapter, the BERT model is designed to generate a language model, thus it only consists of an encoder. BERT can solve most of the common NLP tasks and even performs better than previous NLP models. Apply BERT in various NLP tasks, it only requires adding additional output layers and performing fine-tuning on it.

As the name suggested, BERT learns context from both directions of sequential data (left to right and right to left). BERT model is pre-trained with 2 NLP tasks: Masked Language Modelling (MLM) and Next Sentence Prediction. MLM training is to mask or also known as hiding a word in a sentence, then predict the word based on the context of a sentence. Next Sentence Prediction is to train the BERT model by feeding two sentence inputs and learning to determine if the second sentence input correlates to the first sentence input or is just a simple random sentence. In addition, BERT is pre-trained with multilingual data which is 100 languages.

In this study, a pre-trained BERT model called bert-base-chinese is selected to establish a depression detection model. It is a pre-trained model with simplified and traditional Chinese text data, consisting of 12-layer, 768-hidden nodes, 12-heads and 110 million parameters. It is a huge model. However, it only required fine tuning on the output layer to get a state-of-the-art result.

Before performing training on the model, the dataset will be split into training and testing datasets. The training dataset will be used to train the model, while the testing dataset is to evaluate the performance of the model. Data in the dataset will be split randomly to minimize the biasness.

Then, the model will be trained with a training dataset until it is believed to achieve the best accuracy.

2.3 Evaluating Model

After the model is trained, the model will evaluate with the testing dataset to find out if is there any biasness, underfitting and overfitting in the model. If the result is not as good as the training dataset, it could have biasness within the dataset, or the algorithm itself. Further fine-tuning if needed.

The following show the metrics used to evaluate the performance of the model:

i. Confusion Metrics: It is a performance measure for classification problems.
- True Positive: Predicted positive and it is true.
- False Positive: Predicted positive and it is false.
- False Negative: Predicted negative and it is true.
- True Negative: Predicted negative and it is false.

Table 1
Confusion metrics

	Actual True	Actual False
Predicted True	True Positive (TN)	False Positive (FP)
Predicted False	False Negative (FN)	True Negative (TN)

ii. Accuracy Eq. (1): To calculate how many correct predictions are in percentage.

$$\frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

iii. Precision Eq. (2): To calculate how many correct positive predictions are among the total positive predictions.

$$\frac{TP}{TP+FP} \tag{2}$$

iv. Recall Eq. (3): To calculate how many correct positive predictions are among the total of correct predictions.

$$\frac{TP}{TP+FN} \tag{3}$$

v. F1-score Eq. (4): To measure the balance between precision and recall.

$$\frac{2*Precision*Recall}{Precision+Recall} \tag{4}$$

Besides confusion metrics, the top 20 frequent words will be extracted from both the original testing dataset and the best-trained model prediction. Frequent words will be visualized by word cloud and displayed in tabular form. By side by side comparing the frequent words, allows the researcher to find out which words have higher weightage in the model to make a classification and also the model has a similar depressed word weightage with the testing dataset.

2.4 Fine-Tuning

The following hyperparameters are used in model fine-tuning:

i. Train-Evaluation-Test split ratio: The split ratio of training, evaluating and testing datasets are 60:20:20. Figure 1 shows the distribution for each dataset in term of the number of rows of data.

- Training dataset: 1276695 rows
- Evaluating dataset: 425565 rows
- Testing dataset: 425565 rows

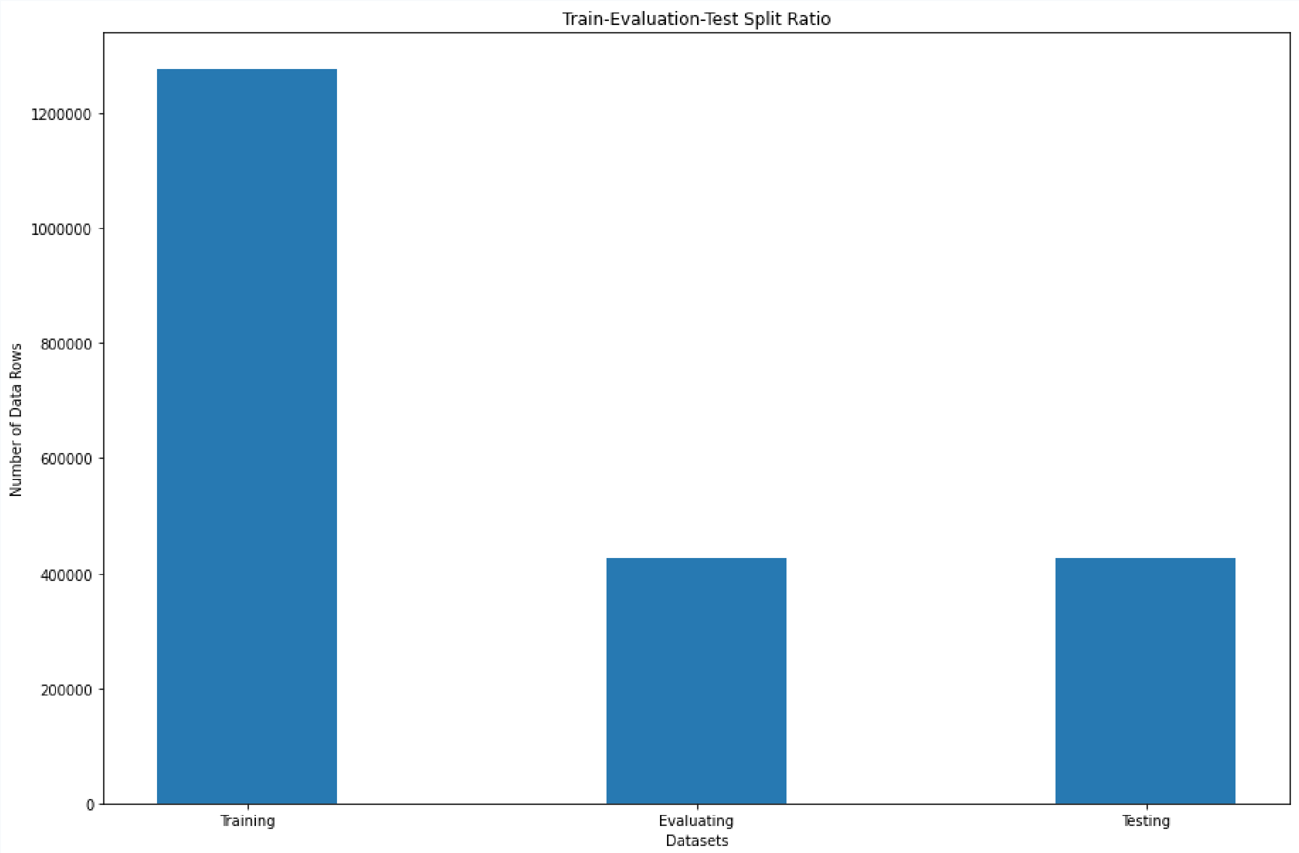

Fig. 1. Train-Evaluation-Test split ratio

ii. Learning rate: According to the BERT author, the learning rates recommended are 3e-4, 1e-4, 5e-5 and 3e-5. Initially, 5e-5 is selected for learning rate during fine-tuning the model. However, the model does not behave expectedly during evaluation as it predicts all data into a single class which is normal. At first, it is believed that this issue is caused by an imbalanced dataset as data labelled as normal has significantly more than depressed data, so the data sampling technique is adopted to balance the ratio of normal and depressed data, which is a ratio of 1:1. After a few iterations of model training and evaluation, this issue does not resolve expectedly. As it is known that a balanced dataset does not help in resolving this issue, it is suggested that it could be because of inappropriate configuration. After changing the learning rate to a smaller figure which is 2e-5, the model is capable of classifying data into both labels. In addition, training and evaluation loss is decreasing gradually throughout the entire training process.

iii. Evaluation strategy: The evaluation strategy refers to when or where to evaluate the training process. In this research, evaluation is conducted every 5000 steps of the training process.

iv. Evaluation steps: Evaluation steps refer to evaluations that take place in every defined number of steps. As mentioned in the evaluation strategy, evaluation is conducted every 5000 steps of the training process.

v. Training batch size: Batch size per Graphics Processing Unit (GPU) for evaluation. In this research, the evaluation batch size is 20. Bigger the batch size, the faster the training process.

vi. Evaluation of batch size: Batch size per Graphics Processing Unit (GPU) for evaluation. In this research, the evaluation batch size is 20. Bigger the batch size, the faster the training process.

vii. Training epoch: The number of training epochs refers to the number of times a model is trained throughout a complete dataset. Since fine-tuning does not require a lot of epochs, 2 epoch is used in this research.

viii. Early stopping: Early stopping refers to a strategy to stop the training when a model's performance on the evaluation dataset starts to degrade. The advantages of early stopping are to avoid over-training on a model which might lead to overfitting and also to save training time. In this research, the model being stop training when the evaluation loss is increasing in three successive times.

2.5 Best Model Selection Strategy

To determine the best model throughout the training, evaluation loss will be used. The training checkpoint with the lowest evaluation loss will be indicated as the best model.

2.6 Data Sampling

As it is an imbalanced dataset (depressed: 408797, normal: 1783113) (Figure 2), the model will be trained with the original dataset, under-sampled dataset and over-sampled dataset, then determine which model performs the best.

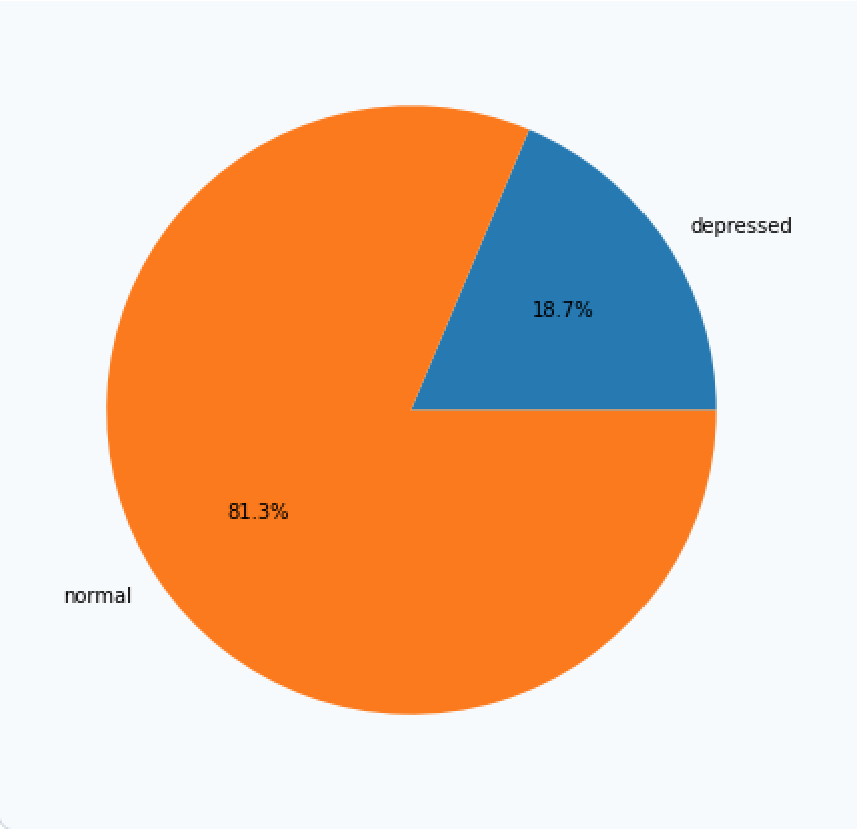

Fig. 2. Distribution of "depressed" and "normal" class

The over-sampling techniques adopted are back-translation and contextual word embeddings. In this research, minority data which is depressed data will be oversampled. Back translation is a text augmentation technique to translate text data into another language and then translate it back to the original language. The contextual word embeddings technique is to either replace or insert a new word in a text sentence.

Over-sampling Example:

i. Back-translation (Mandarin > English > Mandarin) (Table 2):

Table 2
Back-translation

Input	我觉得活着好累
Back-translation	I feel so tired
Result	我觉得好累

ii. Contextual Word Embedding (replace) (Table 3):

Table 3
Contextual word embedding

Input	我觉得活着好辛苦
Result	我觉得活着好*难受*

iii. Contextual Word Embedding (insert) (Table 4):

Table 4
Contextual word embedding

Input	我觉得活着好辛苦
Result	我觉得*生*活着好辛苦

Under-sampling is relatively simple. Data from the majority class which is labelled “normal" will be abundant randomly to create equal-sized classes.

The model will be trained from stretch again with the sampled dataset (over-sampled or under-sampled). Then, compare the performance among three models (trained with the original dataset and sampled dataset) to determine the best model.

3. Research Findings and Discussion

The study aims to explore the capabilities of the BERT model in detecting depression tendencies from Mandarin text sentences. Three pre-trained BERT models are trained according to the original, under-sampled and over-sampled datasets. To determine the best-performed model, each model will be evaluated by a confusion matrix such as accuracy, precision, recall and F1 score.

Besides, visualization technique such as word cloud is presented to the most frequent words that appear in depressing data. By exploring the most frequent words, it could achieve the following objectives. First, to determine if the trained model can predict the depressed sentences correctly. Second, is to determine what features have been learnt by the model.

3.1 Comparison of Performance of Models Trained with Different Sampling Techniques

Table 5 show the performance metrics for each model trained with different data sampling technique. As mentioned before, there are three models trained accordingly. To find out the best-performed model, a confusion matrix will be adopted for evaluation. The table below shows the matrix scored by each model.

Table 5
Model performance comparison

Metrics		Model		
		Original	Under-sampled	Over-sampled
Confusion Matrix	True Positives	335243	295416	329351
	False Positives	9576	49403	15468
	False Negatives	39438	27080	39551
	True Negatives	41309	53667	41196
Accuracy		0.88	0.82	0.87
Precision		0.81	0.52	0.73
Recall		0.51	0.66	0.51
F1-score		0.63	0.58	0.60

From Table 5, the model trained with the original dataset outperforms while compared with others. It is interesting to report that an unbalanced dataset does not affect the performance of the model in classifying normal and depressed data in this research.

The model trained with the under-sampling technique performed the poorest compared with the other model. As data of the majority class is being downsized from 1034446 to 242249, potential important features and information might be lost. Besides, the sampled dataset could lead to the issue of biases. Due to these reasons, the model does not perform well as it is not able to learn comprehensively from the sampled dataset. However, the model scored the highest recall, which shows that the model is more capable of classifying positive samples which is depressing data. It provided an interesting perspective and suggested that the sampled data helps the model narrow down and focus on potential features to make an accurate classification of the positive class although some potential information had been discarded.

The model trained with over-sampling does not perform well as expectedly as well, but it is close to the best-performed model. After performing text augmentation, 792197 data are created and appended to the minority class. However, augmented data are based on the existing dataset, the model might learn new information and knowledge from it. In addition, oversampling could be overfitting as well, but this issue has not been found in this research yet. According to the performance metrics, the model does not benefit from the oversampled dataset and eventually slightly performed poorer than the model trained with the original dataset.

Another evaluation metric called the Receiver Operator Characteristic (ROC) curve is used to measure the performance of binary classification. The curve plots True Positive Rate (TPR) against False Positive Rate (FPR) at different thresholds. An ideal binary classifier will achieve 100% of TPR and 0% of FPR, in which a ROC curve will be presented at a 90% angle to the top-left corner.

In sum, a curve closer to the top-left corner indicates that a model performed better. Area Under the Curve (AUC) measures the area underneath the curve. AUC summarize the ability of a model in distinguishing between 2 classes. The higher the AUC score, a model has better performance in classifying positive and negative classes.

Figure 3 shows the ROC curve and AUC metrics of three trained models. All of the models show a similar curve pattern. However, the model trained with the original dataset scored slightly better than the others.

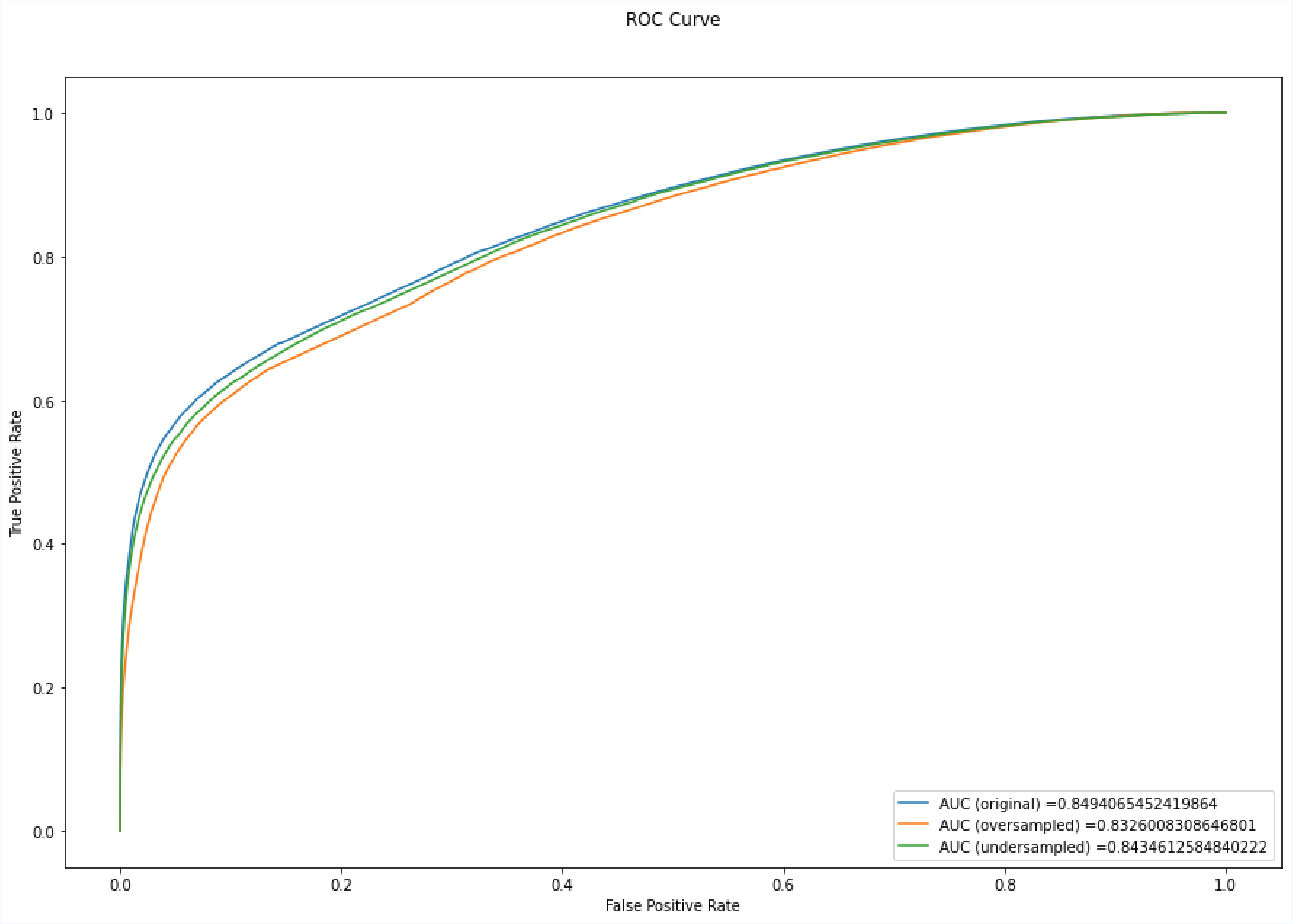

Fig. 3. ROC curve

3.2 Word Cloud and Frequent Words of Depressed Data

Word cloud and word count table are presented in this section to show the most frequent words that are appearing in data which is labelled as depressed. Two-word clouds are presented, one is generated from the original dataset, second is generated based on the model classification. By looking at the word clouds, both word clouds show quite a similar pattern in terms of word sizes. For instance, “想”(think) and “说”(say) are the largest size inside the word cloud (Figure 4 and Figure 5).

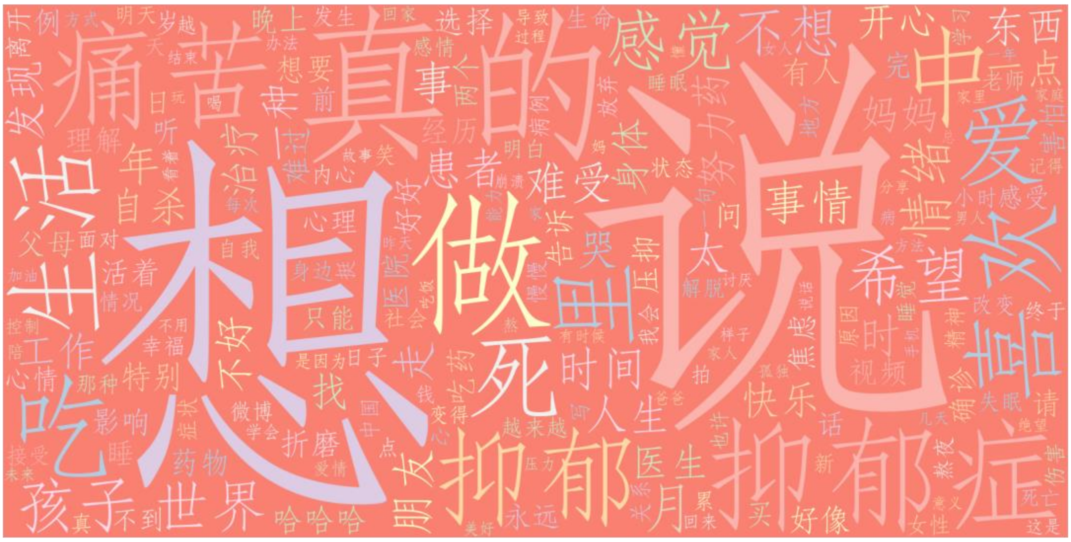

Fig. 4. Word Cloud from testing dataset

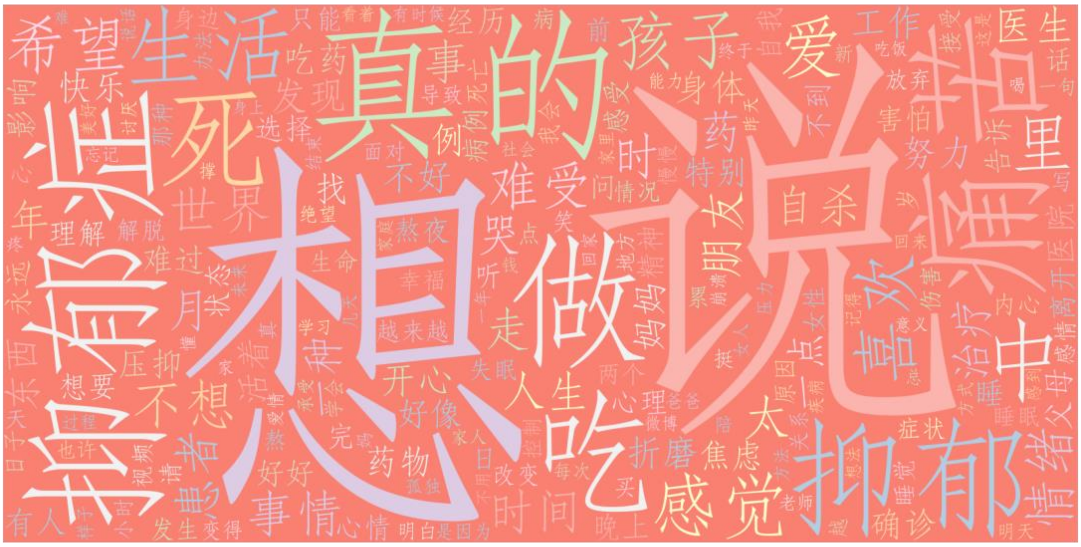

Fig. 5. Word Cloud from the trained model

Table 6 shows the top 20 most frequent words of depressed class data from both the original testing dataset and model classification. This result does not represent the whole knowledge that the model had learned, but it is good to compare if the model could determine the keyword that makes a sentence tend to be in the meaning of depressed.

Despite the natural words like “说” (says), “想” (think) and “真的” (really) which are placed in the top three, both word cloud shows a very similar result. From this result, it can be said that the model had learnt most of the depressing tendency words from the training dataset. For example, words containing negative meanings like “抑郁” (depressed), “抑郁症” (depression), “痛苦” (pain), “死” (die / death) and so on appeared on both tables.

Table 6 also shows a very interesting, phenomenal. Apart from negative words, it is surprising that the word “孩子” (child) also appears very frequent in depressed class data. It suggested that a child could be one of a reason that induces depression emotion. Words “吃” (eat) could lead to medicine intake or anorexia; “希望” (hope) could also suggest to a depressed person that they are hopeless or looking for hope.

Word “情绪” (emotion) replaced “朋友” (friend) in the top 20 from the trained model result. It means that the model learned the word “emotion” tent to have appeared in most of the depressing tendency sentences instead of the word “friend”. Finally, the word “走” (walk) is polysemous. It can mean walking or passing away.

Table 6
Depression words count comparison

Original Testing Dataset		Classification from Trained Model	
Word	Count	Word	Count
说 (say)	15475	说 (say)	13265
想 (think)	10011	想 (think)	8750
真的 (really)	9413	真的 (really)	8407
做 (do / make)	7797	痛苦 (pain)	7627
痛苦 (pain)	7123	做 (do / make)	6488
喜欢 (like / interested)	6219	抑郁症 (depression)	5532
爱 (love)	5873	爱 (love)	5072
抑郁症 (depression)	5430	感觉 (feeling)	4962
感觉 (feeling)	5323	喜欢 (like / interested)	4692
生活 (life)	5295	生活 (life)	4610
希望 (hope)	5030	吃 (eat)	4429
孩子 (child)	4930	孩子 (child)	4252
吃 (eat)	4812	抑郁 (depress)	4212
世界 (world)	4442	希望(hope)	4184
抑郁 (depress)	4067	难受 (suffer)	4024
事情 (happens)	3997	世界 (world)	3871
难受 (suffer)	3700	事情 (happens)	3591
走 (walk)	3575	死 (die / dead)	3311
死 (die / dead)	3380	走 (walk)	3225
朋友(friend)	3205	情绪 (emotion)	2992

From the visualization of word clouds and word counts table, depression data classified by the trained model is similar to the testing dataset. Hence, it can conclude that the model is capable of classifying depression tendency data. In addition, looking at the word cloud and table, also explained what words and features had been learnt by the model to classify depression. For example, the words “痛苦”(pain), “抑郁症”(depression), “抑郁”(depressed) and so on are learnt by the model that most probably appears in depressed data.

Figure 6 shows the line chart of depressed word frequency of both the testing dataset and model prediction.

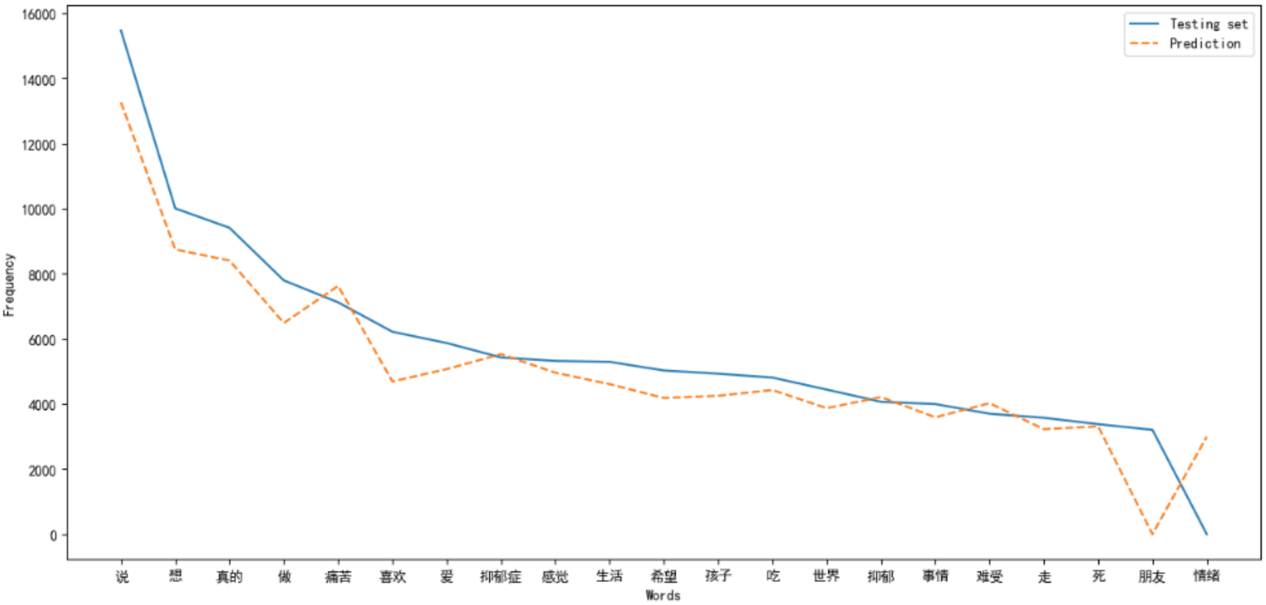

Fig. 6. Line chart of depressed words frequency

4. Conclusions

Depression is one of the most common mental disorders nowadays, especially during the COVID-19 pandemic. Social media is not only a platform for communication and relationship connectivity but also a virtual place to seek for emotional support especially people who are suffering from depression. From the literature review, classification algorithms and machine learning techniques had been studied and experimented with in detecting depression tendencies from the textual content of social media posts. However, most of the existing studies conducted focus on English language text content and Mandarin is relatively lesser. Besides, the attention mechanism-based model is a state-of-the-art model in the NLP domain. Therefore, it is worth finding out the capability of an attention mechanism-based model like BERT in the task of Mandarin textual content depression detection.

In this thesis, the BERT model is adopted and experimented with for depression detection on Mandarin textual data. The model is trained upon the dataset WU3D contains millions of Mandarin textual post data collected from Sina Weibo in China. As the dataset is highly imbalanced, the BERT model will be trained separately by the original dataset, under-sampled dataset and over-sampled dataset. After that, confusion metrics and ROC curve is used to compare the performance among these three models.

From the result, the sampling technique does not help in improving the performance of the model. As discussed in the previous section, significant data might be eliminated through the under-sampling technique. On the other hand, the model does not gain new knowledge and information from the over-sampled dataset because new data is augmented from the existing dataset, which could lead to the issue of overfitting as well.

In conclusion, the BERT model trained by the original dataset performs the best although the dataset is highly imbalanced. It can also conclude that the BERT model is adaptive and work well in the imbalanced dataset. Besides that, this model is also able to detect depressed tendencies in Mandarin text sentences. This study also believed that it could benefit related domains like social care and psychology by early detection of depression on social media and providing timely help.

5. Future Works

This study only focuses on Mandarin language text data. However, the dataset being adopted is from China, it could not represent the entire Chinese community as a whole. For instance, Taiwanese tend to use traditional Chinese; Cantonese communities adopt different types of written systems; Chinese communities from non-Greater China regions such as Singapore and Malaysia might mix local languages in their daily communication. Thus, it is worth exploring and extending the capability of the BERT model in detecting and classifying various Mandarin variants or even mixing them with other languages.

Besides, grammar and word preferences in Malaysian Chinese also have many differences when compared to Chinese from mainland China. For example, Malaysian Chinese tends to not follow the proper Mandarin grammar like the following sentence structure: subject-verb or subject-verb-object. Verbs at the end of a sentence could happen among the Malaysian Chinese community. Due to this reason, the current model might not work well when it has been adopted among the Malaysian Chinese community. To make the model well fit Malaysian Chinese culture, the dataset needs to be collected from sources that originate from Malaysia. The contribution of the dataset could be helpful and benefit future research regardless of the domain of computer sciences, social work and psychology. Meanwhile, we can adopt the model to produce a fun learning as described in the work [32].

Acknowledgment

The funding for this project is made possible through the faculty of Computer Science and Information Technology, UNIMAS.

References

[1] World Health Organization, "Depressive disorder (depression)" (2023). https://www.who.int/news-room/fact-sheets/detail/depression/?gad_source=1&gclid=Cj0KCQjwgrO4BhC2ARIsAKQ7zUmH2UnH2NgKw3LTvq_Duv5YbgrWmcxba3LM0RWq8oBmuAs-MEjBB6saAo6NEALw_wcB

[2] Nazari, Babak, Saeedeh Bakhshi, Marziyeh Kaboudi, Fateme Dehghan, Arash Ziapour, and Nafiseh Montazeri. "A comparison of quality of life, anxiety and depression in children with cancer and healthy children, Kermanshah-Iran." *International Journal of Pediatrics* 5, no. 7 (2017): 5305-5314.

[3] Shadrina, Maria, Elena A. Bondarenko, and Petr A. Slominsky. "Genetics factors in major depression disease." *Frontiers in psychiatry* 9 (2018): 334. https://doi.org/10.3389/fpsyt.2018.00334

[4] Sahimi, Hajar Mohd Salleh, Tuti Iryani Mohd Daud, Lai Fong Chan, Shamsul Azhar Shah, Farynna Hana Ab Rahman, and Nik Ruzyanei Nik Jaafar. "Depression and suicidal ideation in a sample of Malaysian healthcare workers: a preliminary study during the COVID-19 pandemic." *Frontiers in psychiatry* 12 (2021): 658174. https://doi.org/10.3389/fpsyt.2021.658174

[5] Leong Bin Abdullah, Mohammad Farris Iman, Hazwani Ahmad Yusof, Noorsuzana Mohd Shariff, Rohayu Hami, Noor Farahiya Nisman, and Kim Sooi Law. "Depression and anxiety in the Malaysian urban population and their association with demographic characteristics, quality of life, and the emergence of the COVID-19 pandemic." *Current Psychology* (2021): 1-12. https://doi.org/10.1007/s12144-021-01492-2

[6] W. Akram. "A Study on Positive and Negative Effects of Social Media on Society," *International Journal Of Computer Sciences And Engineering*, (2018): 347-354. https://doi.org/10.26438/ijcse/v5i10.351354

[7] Keles, Betul, Niall McCrae, and Annmarie Grealish. "A systematic review: the influence of social media on depression, anxiety and psychological distress in adolescents." *International journal of adolescence and youth* 25, no. 1 (2020): 79-93. https://doi.org/10.1080/02673843.2019.1590851

[8] Orabi, Ahmed Husseini, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. "Deep learning for depression detection of twitter users." In *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*, pp. 88-97. 2018. https://doi.org/10.18653/v1/W18-0609

[9] Wang, Xinyu, Chunhong Zhang, Yang Ji, Li Sun, Leijia Wu, and Zhana Bao. "A depression detection model based on sentiment analysis in micro-blog social network." In *Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2013 International Workshops: DMApps, DANTH, QIMIE, BDM, CDA, CloudSD, Gold Coast, QLD,*

Australia, April 14-17, 2013, Revised Selected Papers 17, pp. 201-213. Springer Berlin Heidelberg, 2013. https://doi.org/10.1007/978-3-642-40319-4_18

[10] Islam, Md Rafiqul, Muhammad Ashad Kabir, Ashir Ahmed, Abu Raihan M. Kamal, Hua Wang, and Anwaar Ulhaq. "Depression detection from social network data using machine learning techniques." *Health information science and systems* 6 (2018): 1-12. https://doi.org/10.1007/s13755-018-0046-0

[11] Yu, Lixia, Wanyue Jiang, Zhihong Ren, Sheng Xu, Lin Zhang, and Xiangen Hu. "Detecting changes in attitudes toward depression on Chinese social media: A text analysis." *Journal of affective disorders* 280 (2021): 354-363. https://doi.org/10.1016/j.jad.2020.11.040

[12] Islam, Md Rafiqul, Muhammad Ashad Kabir, Ashir Ahmed, Abu Raihan M. Kamal, Hua Wang, and Anwaar Ulhaq. "Depression detection from social network data using machine learning techniques." *Health information science and systems* 6 (2018): 1-12. https://doi.org/10.1007/s13755-018-0046-0

[13] Chiong, Raymond, Gregorius Satia Budhi, Sandeep Dhakal, and Fabian Chiong. "A textual-based featuring approach for depression detection using machine learning classifiers and social media texts." *Computers in Biology and Medicine* 135 (2021): 104499. https://doi.org/10.1016/j.compbiomed.2021.104499

[14] Orabi, Ahmed Husseini, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. "Deep learning for depression detection of twitter users." In *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*, pp. 88-97. 2018. https://doi.org/10.18653/v1/W18-0609

[15] Wang, Yiding, Zhenyi Wang, Chenghao Li, Yilin Zhang, and Haizhou Wang. "A multimodal feature fusion-based method for individual depression detection on sina weibo." In *2020 IEEE 39th International Performance Computing and Communications Conference (IPCCC)*, pp. 1-8. IEEE, 2020. https://doi.org/10.1109/IPCCC50635.2020.9391501

[16] Xu, Kaiwei, and Yuhang Fei. "Depression tendency detection of Chinese texts in social media data based on Convolutional Neural Networks and Recurrent neural networks." (2022).

[17] Reece, Andrew G., and Christopher M. Danforth. "Instagram photos reveal predictive markers of depression." *EPJ Data Science* 6, no. 1 (2017): 15. https://doi.org/10.1140/epjds/s13688-017-0110-z

[18] Shen, Guangyao, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, and Wenwu Zhu. "Depression detection via harvesting social media: A multimodal dictionary learning solution." In *IJCAI*, pp. 3838-3844. 2017. https://doi.org/10.24963/ijcai.2017/536

[19] Chowdhary, KR1442, and K. R. Chowdhary. "Natural language processing." *Fundamentals of artificial intelligence* (2020): 603-649. https://doi.org/10.1007/978-81-322-3972-7_19

[20] Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors." *nature* 323, no. 6088 (1986): 533-536. https://doi.org/10.1038/323533a0

[21] Yin, Wenpeng, Katharina Kann, Mo Yu, and Hinrich Schütze. "Comparative study of CNN and RNN for natural language processing." *arXiv preprint arXiv:1702.01923* (2017).

[22] Hochreiter, S. "Long Short-term Memory." *Neural Computation MIT-Press* (1997). https://doi.org/10.1162/neco.1997.9.8.1735

[23] Sutskever, I. "Sequence to Sequence Learning with Neural Networks." *arXiv preprint arXiv:1409.3215* (2014).

[24] Bahdanau, Dzmitry. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* (2014).

[25] Vaswani, A. "Attention is all you need." *Advances in Neural Information Processing Systems* (2017).

[26] Han, Kai, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. "Transformer in transformer." *Advances in neural information processing systems* 34 (2021): 15908-15919.

[27] Jaderberg, Max, Karen Simonyan, and Andrew Zisserman. "Spatial transformer networks." *Advances in neural information processing systems* 28 (2015).

[28] Parmar, Niki, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. "Image Transformer (Proceedings of Machine Learning Research, Vol. 80)." (2018): 4055-4064.

[29] Arnab, Anurag, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. "Vivit: A video vision transformer." In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6836-6846. 2021. https://doi.org/10.1109/ICCV48922.2021.00676

[30] Neimark, Daniel, Omri Bar, Maya Zohar, and Dotan Asselmann. "Video transformer network." In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3163-3172. 2021. https://doi.org/10.1109/ICCVW54120.2021.00355

[31] Devlin, Jacob. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

[32] Abd Kadir, Kauthar, Nurrodhiah Idris, and Alya Izzati Anuar. "Exploring Surah Ad-Dhuha: A Fun and Educational Tafsir App for Kids." *International Journal of Advanced Research in Future Ready Learning and Education* 34, no. 1 (2024): 31-52. https://doi.org/10.37934/frle.34.1.3152