JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY
https://e-journal.uum.edu.my/index.php/jict

How to cite this article:
Sutranggono, A. N., Sarno, R., & Ghozali, I. (2024). Multi-class multi-level classification of mental health disorders based on textual data from social media. *Journal of Information and Communication Technology, 23*(1), 77-104. https://doi.org/10.32890/jict2024.23.1.4

Multi-Class Multi-Level Classification of Mental Health Disorders Based on Textual Data from Social Media

[1]Abi Nizar Sutranggono, *[2]Riyanarto Sarno
& [3]Imam Ghozali
Department of Informatics,
Institut Teknologi Sepuluh Nopember, Indonesia

[1]abi.18052@mhs.unesa.ac.id
*[2]riyanarto@if.its.ac.id
[3]6025211021@mhs.its.ac.id
*Corresponding author

Received: 5/2/2023 Revised: 28/11/2023 Accepted: 10/1/2024 Published: 30/1/2024

ABSTRACT

Mental health disorders pose a significant global public health challenge. Social media data provides insights into these conditions. Analysing text can help identify indications of mental health disorders through text-based analysis. However, despite the large number of studies on the analysis of mental health disorders, the predominant algorithm in the existing literature is the Multi-Class Single-Level (MCSL) classification algorithm, which is often used for simple classification tasks involving a limited number of classes. Typically, these classes are binary, representing either an unhealthy or a healthy mental state. This paper uses English text data from Reddit to classify mental health disorders. The Multi-Class Multi-Level (MCML) classification algorithm was applied to perform detailed classification and address the limitations of the research scope using several approaches, including machine learning, deep learning, and transfer learning approaches. Two different pre-processing scenarios were proposed to handle unstructured text data, one of the most challenging aspects of classifying text from social media. The results of the experiments show that the MCML classification algorithm successfully performs detailed classification and produces promising results for each classification level. The proposed pre-processing scenario influences the performance of each classifier and improves classification accuracy. The best accuracy results were obtained for the Robustly Optimised BERT Pre-training Approach (RoBERTa) classifier at level 1 and level 2 classifications, namely 0.98 and 0.85, respectively. Overall, the MCML classification algorithm is proven to be used as a benchmark for early detection of text-based mental health disorders.

Keywords: MCML classification, mental health disorders, Reddit, text mining, transfer learning.

INTRODUCTION

Mental health disorders, also referred to as mental illnesses (Rehm & Shield, 2019), have been considered one of the world's most severe and pervasive public health problems. Attention-Deficit Hyperactivity Disorder (ADHD), anxiety disorder, Post-Traumatic Stress Disorder (PTSD), bipolar disorder, and other mental illnesses can harm an individual's fitness and well-being. Furthermore, mental health disorders can significantly influence a person's daily life and even lead to suicide (Uban et al., 2021). Nevertheless, individuals suffering from mental disorders often hesitate to consult a qualified specialist when seeking help in treating their illness. Therefore, it is critical to recognise early signs of mental health disorders exhibited by a certain individual so that their development can be monitored. In some cases, this can have a significant impact, even saving someone's life (Zhang et al., 2022). On the other hand, data obtained from social media platforms can be used to gain a deeper understanding of individual behaviour, mental health conditions, and developments (Ramírez-Cifuentes et al., 2020). Real-time information on social media is abundant, constantly changing, and relatively easy to collect. Nowadays, people have shared their interests, daily activities, moods, and feelings on their social media platforms.

Numerous studies have discussed the automatic detection of mental illnesses using social media data with varying focuses and methods, most of which used Natural Language Processing (NLP) approaches. Previous studies have primarily focused on understanding depression (Eichstaedt et al., 2018; Ríssola et al., 2019; Tadesse et al., 2019) and other mental illnesses such as anxiety disorder, schizophrenia, and PTSD (Bae et al., 2021; Chang & Tseng, 2020; Coppersmith et al., 2014). Various social media platforms, including Twitter, Facebook, and Reddit, have been considered. Additionally, most studies employed a classical machine learning approach to perform quantitative analysis or develop classification models (Kim et al., 2021), while several other studies have utilised more sophisticated deep learning techniques, i.e., Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and many others (Iyortsuun et al., 2023; Kim et al., 2020; Su et al., 2020). Furthermore, transfer learning approaches have been increasingly employed in current text-mining studies as researchers strive to improve accuracy and performance by applying various transformers. In this regard, one study has explored several methods, including Bidirectional Encoder Representations from Transformers (BERT), XLNet, and Robustly Optimised BERT Pre-training Approach (RoBERTa) transfer learning approaches, for classifying mental illnesses based on textual data from the social media Reddit (Ameer et al., 2022). RoBERTa is a transformer-based model that outperforms BERT in various benchmark tasks. However, studies of text-based mental health analysis using a transfer learning-based approach are still very limited and need to be explored further.

Although a number of studies have addressed the topic of mental health disorder analysis, the focus has been limited to simple classification problems (Ríssola et al., 2022). The Multi-Class Single-Level (MCSL) classification algorithm is commonly used in the existing literature to classify a limited number of classes at a single level (Bae et al., 2021; Chang & Tseng, 2020; Coppersmith et al., 2014; Tadesse et al., 2019). These classes are typically binary, representing either an unhealthy or a healthy mental state. Meanwhile, machine learning approaches can solve not only simple classification problems but also Multi-Class Multi-Level (MCML) classification problems. Additionally, the MCML classification algorithm can be used to perform detailed classification and obtain better classification performance. This algorithm has been adapted in a prior study whose subject is image classification (Hameed et al., 2020). In other words, no studies have implemented the concept of MCML text classification. To overcome existing research limitations, this study conducted a detailed text-based classification of mental illnesses on Reddit English text data using the MCML classification algorithm in several approaches, namely machine learning, deep learning, and transfer learning.

Social media posts frequently include emojis, abbreviations, and unique phrases. Hence, texts originating from social media are usually unstructured, making it challenging to analyse mental health disorders using social media textual data. Thus, an appropriate method is highly needed to handle unstructured text data from social media to obtain valid analytical results for detecting mental illnesses. A study by HaCohen-Kerner et al. (2020) addressed the impacts of pre-processing on text classification with an emphasis on the pivotal role of pre-processing techniques in enhancing classification accuracy. In addition, a review article by Zhang et al. (2022) underscored the significance of NLP techniques for detecting mental illnesses from text, highlighting the necessity for reliable pre-processing methods to support early detection, prevention, and treatment of mental disorders. In this present study, on the other hand, a modified pre-processing technique based on the characteristics of the dataset was implemented to handle unstructured social media textual data. After the Introduction section, this paper is organised as follows where the Related Work section describes relevant theories while the Proposed Method section details the deployed procedures. On the other hand, the Results section presents the outcomes of the experiments, and the conclusion of the work is provided in the last section.

RELATED WORK

This section is divided into several subsections, explaining several areas pertinent to this study. Concepts regarding mental health disorders are discussed thoroughly in the first subsection, while the MCML classification algorithm is explored in the following subsection. Finally, an exposition of machine learning, deep learning, and transfer learning approaches is also provided.

Mental Health Disorders

Mental health disorders, also known as mental illnesses, are syndromes characterised by changes in a person's emotions, thought patterns, and behaviour that reflect biological, psychological, and/or mental dysfunctions (Stein et al., 2021). Many people suffer from stress, depression, or anxiety, which can also be considered early symptoms of mental disorders if these conditions have progressed beyond normal limits to the point of interfering with social functioning or resulting in a decline in physical health. Mental illnesses can disrupt daily life and cause a variety of social issues, such as the inability to work with others or maintain relationships. A psychiatrist can provide counselling and therapy to treat these symptoms and mental illnesses. There are various types of mental illnesses, some of which are described as follows. Bipolar disorder is a mental health condition that causes extreme mood swings. Anxiety is characterised by excessive feelings of worry and fear. Depression is a medical condition that has a broad and negative impact on an individual's feelings, thoughts, and actions. PTSD is a mental illness that occurs after a traumatic, frightening, or dangerous event. Lastly, ADHD is a mental illness that affects how people focus their attention and regulate their behaviour.

Multi-Class Multi-Level Classification Algorithm

In their study, Hameed et al. (2020) proposed an intelligent Multi-Class Multi-Level (MCML) classification algorithm inspired by the "divide and conquer" rule that classifies a class into several subclasses at multiple levels, thus allowing for more specific classification and resulting in improved classification performance. The MCML classification algorithm is proven able to solve problems of image data classification. In the field of text mining, the most commonly used algorithm in the existing literature is the Multi-Class Single-Level (MCSL) classification algorithm, which completes simple classification tasks with limited classes. A previous related study by Ameer et al. (2022) has classified datasets with multiple classes, i.e., ADHD, anxiety, bipolar, depression, PTSD, and none, at a single level, as shown in Figure 1. ADHD, anxiety, bipolar, depression, and PTSD can all be classified as mental health disorders or mental illnesses, while the remaining class, 'none', can be referred to as non-mental illness. Therefore, in this study, binary classification was performed at level 1 using the MCML classification algorithm by considering two categories: mental illness and non-mental illness. Texts indicating mental illnesses were further classified at level 2 into ADHD, anxiety, bipolar, depression, and PTSD. Figure 2 illustrates the class division in the implementation of the MCML classification algorithm at levels 1 and 2.

Figure 1

Multi-Class Single-Level Classification

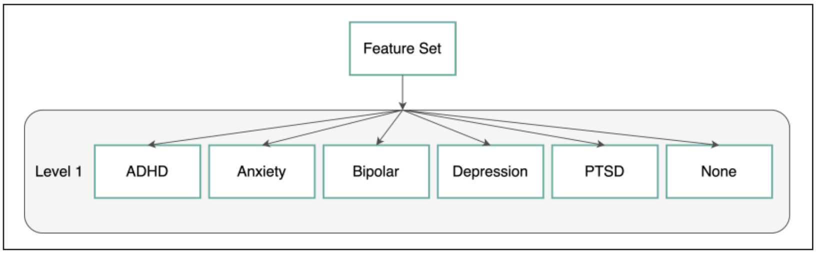
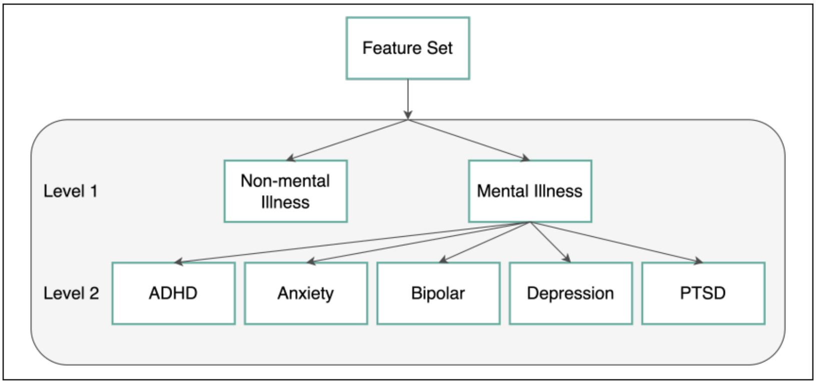


Figure 2

Multi-Class Multi-Level Classification



Machine Learning Approach

Machine learning methods can be applied to text classification. The initial dataset is split into two datasets for training and testing purposes. The training dataset encompasses the data required to train the classifiers for specific data characteristics. The performance of the model developed from the training dataset is then assessed using the testing dataset. The machine learning algorithms applied for text analysis are classified as supervised classification (Noori, 2021). Specifically, this study implemented a machine learning approach with several algorithms, namely Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB), Stochastic Gradient Descent (SGD), and Support Vector Machine (SVM) with linear, RBF, and polynomial kernel functions. LR is a machine learning model that employs a statistical approach to assign discrete labels to data points and locates any decision boundaries that distinguish classes (Hsu, 2020). Meanwhile, RF consists of many decision trees, where voting is carried out on the results of the decision trees; the category with the most votes is the outcome of the RF algorithm (Sun et al., 2020). NB is a well-known simple classifier based on Bayes' theorem, which is widely used in document classification (Hosseini et al., 2022). SGD is a variation of Gradient Descent (GD) that deals with random (stochastic) probability where a single sample is chosen for model training at each iteration (Gaye et al., 2021). Lastly, SVM is a machine learning algorithm that uses a vector approach to classify related documents to generate a linear hyperplane with an optimised margin between target classes (Luo, 2021).

Deep Learning Approach

Deep learning is a branch of machine learning that processes data or signals similarly to that of human intelligence (Sarker, 2021). Multiple layers of thousands of neurons are used in deep learning models to communicate with each other, enabling faster parallel processing. Deep learning algorithms such as RNNs and CNNs can be utilised for text classification. Deep learning models extract features or patterns, thus eliminating the need to construct them manually from text. In this study, a deep learning approach was implemented by creating Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) models. LSTM is a type of RNN that is capable of learning long-term dependencies. This model has a chain-like structure similar to RNNs, but its base module is structurally distinct from other RNNs. While other RNNs are usually plagued by vanishing and exploding gradient problems, which limit their ability to learn and understand long sequences and contexts, LSTMs are not affected by this issue (Al Hamoud et al., 2022; Rai et al., 2022). Meanwhile, GRU has significantly faster performance than LSTM due to its computational simplicity.

Transfer Learning Approach

Another branch of machine learning is transfer learning. As traditional approaches assume that the dataset is within the same scope, a different model is required when the scope changes. The transfer learning method enables classifiers that have been previously trained within a particular scope to be reused in other scopes (Nandwani & Verma, 2021). By employing this method, classifiers trained on massive datasets to overcome one task can be used to solve similar problems. Using pre-trained models as a foundation for a related scope can save time and lead to better outcomes. In this study, the transfer learning approach was used by adopting several pre-trained models built on transformers. In solving various NLP problems, BERT models deliver cutting-edge performance. BERT consists of a transformer attention mechanism that learns contextual relationships between words. The transformer is made up of an encoder that reads text input and a decoder that is in charge of making task-based predictions. The BERT model has been trained on a massive English corpus to learn English text data (Devlin et al., 2018).

Aside from the BERT model, another state-of-the-art model is the RoBERTa model, which improves the BERT model by modifying key hyperparameters and training on larger datasets (Liu et al., 2019). More specifically, the RoBERTa model is proposed to improve the BERT base model by employing large mini-batches, larger Byte-Pair Encoding (BPE) techniques, and dynamic masking, a complete sentence without Next Sentence Prediction (NSP) loss. By combining these techniques, the RoBERTa language model improves performance by 2 percent to 4 percent over the BERT base model in most NLP downstream tasks. In addition to the BERT and RoBERTa models, this study also used other pre-trained models, namely a lite BERT (ALBERT) and a distilled BERT (DistilBERT). ALBERT is an extended version of the BERT model proposed to make BERT more scalable and reduce computational power during training by using two techniques: decomposing the embedding parameters into two smaller matrices and sharing the parameters across multiple layers (Lan et al., 2019). On the other hand, DistilBERT is a small, fast, low-cost transformer model trained with a BERT base (Sanh et al., 2019). This model has 40 percent fewer parameters than BERT-base-uncased and runs 60 percent faster while maintaining 95 percent of BERT's performance.

THE PROPOSED METHOD

The methodology of this study is described in this section. The initial step involves using data pre-processing to refine and organise the dataset in preparation for further analysis. Afterwards, feature engineering methods are adopted to convert textual data into numerical values, a prerequisite for effectively utilising machine learning and deep learning models. Subsequently, the MCML algorithm is employed in the classification process. To assess the efficacy of the proposed method, a comprehensive evaluation of the performance of each classifier is carried out, which provides valuable insights into the accuracy and effectiveness of the model in handling text classification tasks.

Data Acquisition

The data used to build the classification model in this study comes from the Reddit dataset (Amurark, 2021), a collection of English Reddit records obtained through data crawling through the Reddit API. This dataset consists of 16,703 text records. Two types of text data are to be combined for analysis: the title and the content. First, the dataset was divided into three groups of CSV files, namely training, validation, and testing. Every Reddit record was classified into six categories: ADHD, anxiety, bipolar, depression, and PTSD, and none. The distribution of data for each class from all groups is displayed in Figure 3, where the difference in the number of records in each class is not statistically significant. Specifically, Table 1 presents the proportion of data per class for each set. The total data from the training, validation, and testing datasets are 13,727, 1,488, and 1,488, respectively. Furthermore, Table 2 shows the samples of the Reddit dataset. The dataset has already been pre-processed by removing URLs and usernames.

Figure 3

Distribution of Data for Each Class

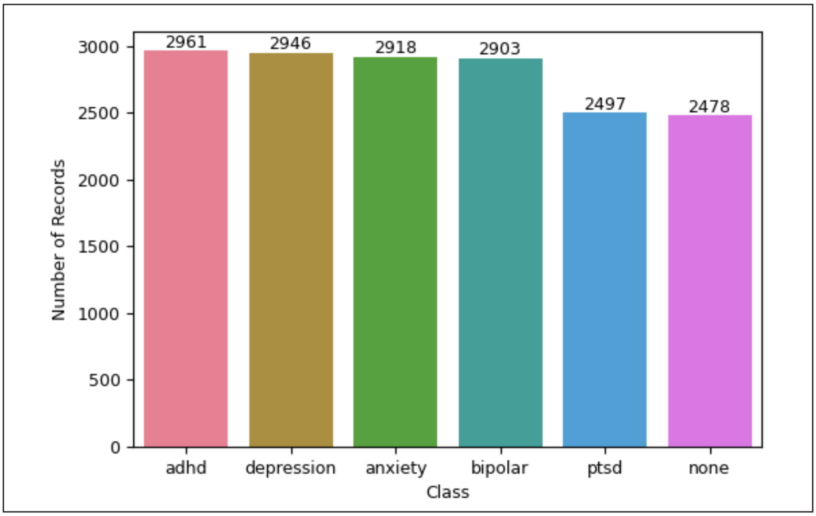


Table 1

Proportion of Data for Each Set

Class	Training data	Validation data	Testing data	Total
ADHD	2465	248	248	2961
Anxiety	2422	248	248	2918
Bipolar	2407	248	248	2903
Depression	2450	248	248	2946
PTSD	2001	248	248	2497
None	1982	248	248	2478
Total	13727	1488	1488	16703

Table 2

Samples of Reddit Dataset

Text	Class
awful rules, pointless bureaucratic redlining bullshit, fierce attention to mind numbing detail, but we have to craft good applications to do the right things first! fucking hell, i hate this disorder.	ADHD
because i do. i miss those days when i used to smile a lot, laughed a lot, and talk a lot. i keep avoiding my past friends now because i'm such a mess. as i was studying at school, all i can do now is to watch my old friends grow up with me at the background. they just grow up so fast.	Anxiety
anyone else? it's been creeping up for a few weeks. i just want to be alone... i actually am alone emotionally but not physically (kids, family). this is shit.	Bipolar
probably sounds stupid but the only thing i look forward to everyday is laying in bed after doing nothing all day and just sleeping for 15 hours. i wish i could sleep longer	Depression
i nearly jumped out of my skin because a coworker simply said good morning to me. i wish i could just be normal instead of being that guy people are told to avoid talking to.	PTSD
when i use "and", can i remove the second possessive adjective? 1. your name and age 2. your name and your age which one sounds more correct? or is there any difference between them (like one is formal and the other is informal)? thank you for the help	None

Data Pre-processing

The identification of mental health disorders using social media data is challenging due to the unstructured textual data derived from social media. In addition, emojis, abbreviations, or unique phrases are commonly found in social media textual data. Consequently, the textual data must be prepared through a pre-processing process before text classification. Text pre-processing attempts to convert unstructured text data into structured data that can be easily analysed (Duong & Nguyen-Thi, 2021).

This study proposed two pre-processing scenarios. Several pre-processing libraries, including Natural Language Toolkit (NLTK), demoji, and RegEx, were utilised in this study. Figure 4 illustrates the pre-processing stage of the dataset using scenario X, while Figure 5 depicts the pre-processing stage of the dataset using scenario Y. In general, the difference between scenarios X and Y is that scenario X transforms emojis and emoticons to corresponding words and uses a spell checker tool to correct spelling, whereas scenario Y removes emojis and emoticons from the text and does not use a spell checker tool. In both scenarios, sensitive information removal was used to remove sensitive information, such as hashtags. The contraction transformation procedure was carried out to normalise word forms with contractions. Lower casing converted each letter in the text to lowercase. After that, each abbreviation was transformed into a common word in the abbreviation transformation step. Punctuations and numbers were then removed from the text during the punctuation removal and number removal processes. When the multi-whitespace removal process took place, excess spaces were removed. Similarly, the single letter in the text was eliminated during the single-letter removal step. Furthermore, stop words were removed during the stop word removal process. The results of the pre-processing stage were saved in a CSV format file. Table 3 presents examples of pre-processing results for scenarios X and Y.

Figure 4

Pre-processing Stage of Scenario X

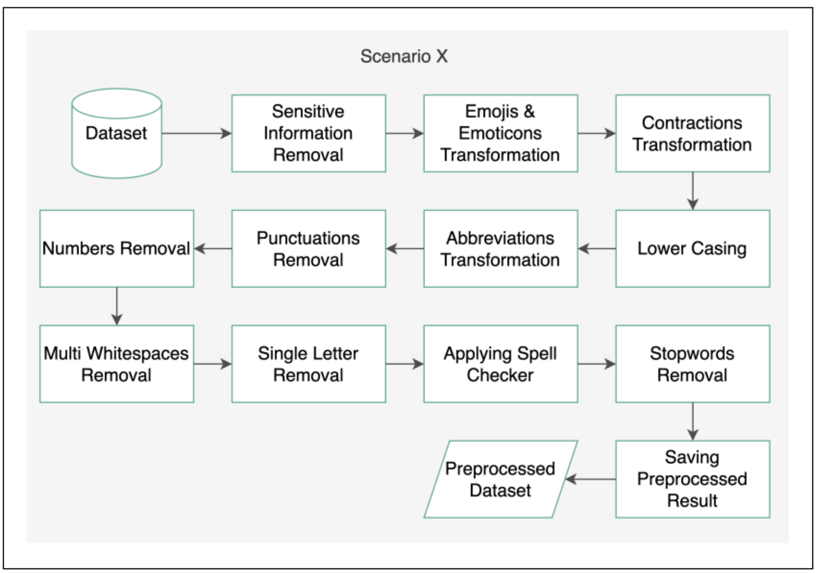


Figure 5

Pre-processing Stage of Scenario Y

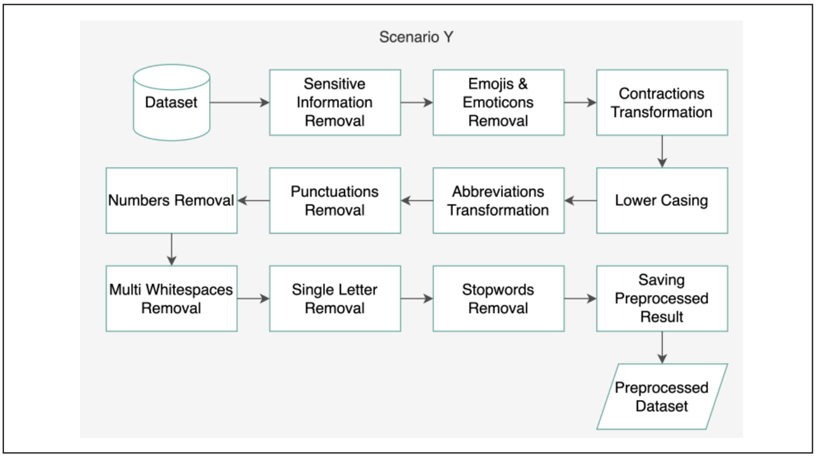


Table 3

Results of Data Pre-processing

Before pre-processing	Pre-processing result of scenario X	Pre-processing result of scenario Y
does anyone else feel like a psycho while having conversation with yourself in your mind? i was just recalling my past and having long conversation with myself in my mind as usual, sarcastically speaking and dark humour made me laugh and i felt like a psycho tbh. edit; thank you, yall! it really gives me relief knowing i'm not alone and that it's normal. i mean, i knew it's normal thing to do but i just couldn't help but doubt myself sometimes.	anyone else feel like psycho conversation mind recalling past long conversation mind usual sarcastically speaking dark humour made laugh felt like psycho honest edit thank really gives relief knowing alone normal mean knew normal thing could help doubt sometimes grinning face sweat	anyone else feel like psycho conversation mind recalling past long conversation mind usual sarcastically speaking dark humour made laugh felt like psycho honest edit thank really gives relief knowing alone normal mean knew normal thing could help doubt sometimes

Feature Engineering

To classify text using a machine learning or deep learning approach, feature engineering methods must be applied to convert text data into numerical values. The Term Frequency-Inverse Document Frequency (TF-IDF) feature engineering method was used in this study to implement the machine learning approach. On the other hand, to implement the deep learning approach, the FastText feature engineering method was also applied in this study. TF-IDF weights each word whose numerical representation will be sought (Taha et al., 2021). The weight assigned to a word by TF-IDF reflects the importance of the word in the document. TF-IDF is one of feature engineering methods that consider word frequency in text (Maryamah et al., 2021). Meanwhile, FastText is an extension of the Word2Vec Embedding method, which employs the concept of similarity of meaning of a word to solve the Out of Vocabulary (OOV) problems (Khattak et al., 2019). This study adopted a 300-dimensional pre-trained FastText model containing 1 million word vectors trained on Wikipedia 2017, the UMBC web-based corpus, and the statmt.org news dataset (Mikolov et al., 2017).

Classification

The MCML algorithm was used during the classification step. This algorithm employs the rule of 'divide and conquer' and divides the classification task into sub-classification tasks. At level 1, binary classification was performed to classify texts into two categories: mental illness and non-mental illness. At level 2, mental illness texts were further classified into ADHD, anxiety, bipolar, depression, and PTSD. In this study, the MCML classification algorithm was applied using machine learning, deep learning, and transfer learning approaches. This study implemented a machine learning-based approach with several algorithms, including LR, RF, NB, SGD, and SVM. In more detail, this study employed SVM with variations of kernel functions. The deep learning approach was implemented by constructing LSTM and GRU models. In addition, several pre-trained models were used to perform the MCML classification using the transfer learning-based approach, namely the BERT base uncased model, the DistilBERT model, the ALBERT model, and the RoBERTa model. Each classifier is built with customised settings and was run multiple times to achieve the best results.

Evaluation

The constructed classifiers were tested using the testing data, and the performance of each classifier was assessed using evaluation metrics. Additionally, the time taken by the classifiers to classify text or the testing time was used as a benchmark for assessing the effectiveness of each classification model. Accuracy, precision, recall, and F1 score were the evaluation metrics used in this study. Accuracy represents the classifier's ability to classify sentiments correctly. Precision is the ratio of the classifier results that are correctly predicted to the requested data information. The proportion of correct predictions of classification results to the amount of actual data information is known as recall. Meanwhile, the F1 score compares the average precision and recall values of the test results. The formulas for calculating accuracy, precision, recall, and F1 score are presented in Equations 1 to 4.

$$Accuracy = \frac{TP+TN}{(TP+FN)+(FP+TN)} \tag{1}$$

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

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

$$F1\ score = \frac{2TP}{2TP+FN+FP} \tag{4}$$

where true positive, true negative, false positive, and false negative are represented by TP, TN, FP, and FN,respectively (Vinod Prakash & Dharmender Kumar, 2023).

RESULTS AND DISCUSSION

The classification task in the MCML classification algorithm was divided into two levels. At level 1, binary classification divided texts into two categories: mental illness and non-mental illness. Mental illness texts were further classified at level 2 into ADHD, anxiety, bipolar, depression, and PTSD. Furthermore, this study used two pre-processing scenarios, i.e., scenarios X and Y.

Table 4 shows the testing time of each machine learning classifier in classifying text for MCML level 1 classification. The testing time of the Polynomial SVM classifier combined with the scenario Y pre-processing technique is the longest among other classifiers when employing the classical machine learning approach. Meanwhile, the classifier with the shortest testing time is the NB classifier combined with the scenario X pre-processing technique. Table 5 displays the evaluation results of the machine learning classifiers for MCML level 1 classification, where the SGD classifier achieves the highest overall accuracy value of 0.9798 in both combinations with the scenario X and scenario Y pre-processing techniques.

Table 4

Testing Time of Machine Learning Classifiers for MCML Level 1 Classification

Classifier	Scenario X	Scenario Y
LR	0.30 s	0.61 s
RF	0.40 s	1.13 s
NB	0.29 s	0.79 s
Linear SVM	55.72 s	170.14 s
RBF SVM	246.63 s	456.36 s
Polynomial SVM	400.47 s	488.15 s
SGD	0.36 s	0.68 s

Table 5

Evaluation of Machine Learning Classifiers for MCML Level 1 Classification

Classifier	Scenario	Accuracy	Precision	Recall	F1 score
LR	X	0.9549	0.9558	0.9549	0.9525
	Y	0.9616	0.9623	0.9616	0.9599
RF	X	0.9274	0.9308	0.9274	0.9196
	Y	0.9254	0.9295	0.9254	0.9169
NB	X	0.8864	0.9000	0.8864	0.8606
	Y	0.8763	0.8923	0.8763	0.8441
Linear SVM	X	0.9731	0.9729	0.9731	0.9725
	Y	0.9778	0.9776	0.9778	0.9775
RBF SVM	X	0.9684	0.9685	0.9684	0.9674
	Y	0.9711	0.9711	0.9711	0.9703
Polynomial SVM	X	0.8413	0.8667	0.8413	0.7763
	Y	0.8413	0.8667	0.8413	0.7763
SGD	X	0.9798	0.9796	0.9798	0.9796
	Y	0.9798	0.9797	0.9798	0.9795

Table 6 compares the testing time of each deep learning classifier in classifying text for MCML level 1 classification. When applying the deep learning approach, the testing time of the LSTM classifier combined with the scenario Y pre-processing technique is the longest among other classifiers. Meanwhile, the GRU classifier combined with the scenario X pre-processing technique has the shortest testing time. Table 7 reveals the evaluation results of the deep learning classifiers for MCML level 1 classification, where the GRU classifier combined with the scenario Y pre-processing technique has the best overall accuracy, precision, recall, and F1 score values of 0.9825. This indicates that GRU is superior to LSTM in detecting mental illnesses.

Table 6

Testing Time of Deep Learning Classifiers for MCML Level 1 Classification

Classifier	Scenario X	Scenario Y
LSTM	50.12 s	56.79 s
GRU	40.36 s	40.51 s

Table 7

Evaluation of Deep Learning Classifiers for MCML Level 1 Classification

Classifier	Scenario	Accuracy	Precision	Recall	F1 score
LSTM	X	0.9677	0.9673	0.9677	0.9674
	Y	0.9724	0.9721	0.9724	0.9719
GRU	X	0.9771	0.9769	0.9771	0.9769
	Y	0.9825	0.9825	0.9825	0.9825

Table 8 displays the comparison of the testing time of each transfer learning classifier in classifying text for MCML level 1 classification. When using the transfer learning approach, the ALBERT classifier combined with the scenario Y pre-processing technique takes 27.63 seconds, which is longer than the other classifiers. Meanwhile, with a testing time of 12.97 seconds, the DistilBERT classifier combined with the scenario Y pre-processing technique has the shortest testing time. On the other hand, the evaluation results of the transfer learning approach for MCML level 1 classification are shown in Table 9. The

DistilBERT and RoBERTa classifiers outperform the other classifiers when combined with the scenario Y pre-processing technique. For both classifiers, the values of the accuracy, recall, precision, and F1 score are all 0.9899.

Table 8

Testing Time of Transfer Learning Classifiers for MCML Level 1 Classification

Classifier	Scenario X	Scenario Y
BERT	25.92 s	25.61 s
ALBERT	24.59 s	27.63 s
DistilBERT	14.93 s	12.97 s
RoBERTa	23.40 s	24.88 s

Table 9

Evaluation of Transfer Learning Classifiers for MCML Level 1 Classification

Classifier	Scenario	Accuracy	Precision	Recall	F1 score
BERT	X	0.9892	0.9892	0.9892	0.9892
	Y	0.9872	0.9872	0.9872	0.9872
ALBERT	X	0.9831	0.9832	0.9831	0.9832
	Y	0.9858	0.9862	0.9858	0.9859
DistilBERT	X	0.9865	0.9866	0.9865	0.9865
	Y	0.9899	0.9899	0.9899	0.9899
RoBERTa	X	0.9831	0.9831	0.9831	0.9831
	Y	0.9899	0.9899	0.9899	0.9899

The testing times of machine learning classifiers for MCML level 2 classification are presented in Table 10. When employing the classical machine learning approach, the testing time of the RBF SVM classifier combined with the scenario Y pre-processing technique is longer than the other classifiers, namely 591.29 seconds. On the other hand, the NB classifier combined with the scenario X pre-processing technique only takes 0.46 seconds, which is the shortest testing time among all classifiers. In addition, Table 11 shows the evaluation results of the machine learning approach for MCML level 2 classification.

Combined with the scenario Y pre-processing technique, the SGD classifier obtains the highest overall accuracy value of 0.8096. Meanwhile, the precision, recall, and F1 score values of this classifier are 0.8161, 0.8096, and 0.8107, respectively.

Table 10

Testing Time of Machine Learning Classifiers for MCML Level 2 Classification

Classifier	Scenario X	Scenario Y
LR	0.62 s	0.64 s
RF	1.14 s	0.86 s
NB	0.46 s	0.54 s
Linear SVM	188.22 s	263.51 s
RBF SVM	440.73 s	591.29 s
Polynomial SVM	181.72 s	346.37 s
SGD	0.77 s	1.14 s

Table 11

Evaluation of Machine Learning Classifiers for MCML Level 2 Classification

Classifier	Scenario	Accuracy	Precision	Recall	F1 score
LR	X	0.7846	0.7916	0.7846	0.7848
	Y	0.8080	0.8145	0.8080	0.8088
RF	X	0.7330	0.7561	0.7330	0.7330
	Y	0.7629	0.7778	0.7629	0.7650
NB	X	0.7177	0.7438	0.7177	0.7137
	Y	0.7459	0.7700	0.7459	0.7449
Linear SVM	X	0.7725	0.7779	0.7725	0.7724
	Y	0.7903	0.7976	0.7903	0.7912
RBF SVM	X	0.7693	0.7807	0.7693	0.7696
	Y	0.7943	0.8060	0.7943	0.7957
Polynomial SVM	X	0.6733	0.7106	0.6733	0.6590
	Y	0.6975	0.7335	0.6975	0.6898
SGD	X	0.7790	0.7827	0.7790	0.7782
	Y	0.8096	0.8161	0.8096	0.8107

Table 12 compares the testing time of each deep learning classifier in classifying text for MCML level 2 classification, where the LSTM classifier combined with the scenario X pre-processing technique has the longest testing time. Meanwhile, when combined with the scenario X pre-processing technique, the GRU classifier has the shortest testing time. The evaluation results of the deep learning approach for MCML level 2 classification are shown in Table 13. When using the deep learning approach, the GRU classifier combined with the scenario Y pre-processing technique produces the best overall results. The accuracy of this classifier is 0.8072, while its precision, recall, and F1 score values are 0.8168, 0.8072, and 0.8073, respectively.

Table 12

Testing Time of Deep Learning Classifiers for MCML Level 2 Classification

Classifier	Scenario X	Scenario Y
LSTM	35.66 s	26.27 s
GRU	22.91 s	23.53 s

Table 13

Evaluation of Deep Learning Classifiers for MCML Level 2 Classification

Classifier	Scenario	Accuracy	Precision	Recall	F1 score
LSTM	X	0.4483	0.4493	0.4483	0.4151
	Y	0.7072	0.7429	0.7072	0.7078
GRU	X	0.7701	0.7808	0.7701	0.7687
	Y	0.8072	0.8168	0.8072	0.8073

Table 14 displays the testing time of each transfer learning classifier in classifying text for MCML level 2 classification. When using the transfer learning approach, the BERT classifier combined with the scenario X pre-processing technique takes 23.39 seconds, which is longer than the other classifiers. Meanwhile, with a testing time of 11.25 seconds, the DistilBERT classifier combined with the scenario Y pre-processing technique has the shortest testing time. Table 15 reveals the evaluation results of the transfer learning approach for MCML level 2 classification. The RoBERTa transfer learning classifier combined with the scenario Y pre-processing technique outperforms the other classifiers with an accuracy value of 0.85, which is quite good for this complex task of mental illness classification. Furthermore, the precision, recall, and F1 score values of the RoBERTa classifier are 0.8524, 0.8500, and 0.8503, respectively.

Table 14

Testing Time of Transfer Learning Classifiers for MCML Level 2 Classification

Classifier	Scenario X	Scenario Y
BERT	23.39 s	21.43 s
ALBERT	20.45 s	23.16 s
DistilBERT	12.38 s	11.25 s
RoBERTa	21.84 s	21.39 s

Table 15

Evaluation of Transfer Learning Classifiers for MCML Level 2 Classification

Classifier	Scenario	Accuracy	Precision	Recall	F1 score
BERT	X	0.8177	0.8192	0.8177	0.8174
	Y	0.8330	0.8369	0.8330	0.8336
ALBERT	X	0.7983	0.8002	0.7983	0.7980
	Y	0.8185	0.8229	0.8185	0.8192
DistilBERT	X	0.8040	0.8098	0.8040	0.8034
	Y	0.8330	0.8363	0.8330	0.8334
RoBERTa	X	0.8161	0.8193	0.8161	0.8154
	Y	0.8500	0.8524	0.8500	0.8503

Based on the overall results of the experiments, the RoBERTa classifier combined with the scenario Y pre-processing technique outperforms other classifiers in solving MCML level 1 and 2 classification tasks. This classifier obtains an accuracy of 0.9899 for MCML level 1 classification and 0.85 for MCML level 2 classification. RoBERTa is an improved variant of BERT that removes the Next Sentence Prediction (NSP) task and concentrates on the pre-training masked language modelling using a dynamic masking pattern. In particular, masked language modelling aids the model's performance and produces a better result. Table 16 presents the average accuracy of each approach applied in this study. Among the three approaches used to implement the MCML classification algorithm, the average accuracy of the transfer learning approach is better than that of machine learning and deep learning approaches. The average accuracy values of the transfer learning approach are 0.9869 for MCML level 1 classification and 0.8213 for MCML level 2 classification. Transfer learning is implemented using pre-trained models built on transformers that rely on the attention mechanism to provide a wider range of contexts in a given sequence, allowing for better classification performance (Casola et al., 2022).

Table 16

Average Accuracy of Each Approach

Classification	Approach	Average accuracy
MCML level 1	Machine learning	0.9331
	Deep learning	0.9749
	Transfer learning	0.9868
MCML level 2	Machine learning	0.7598
	Deep learning	0.6832
	Transfer learning	0.8213

The best accuracy values obtained in this study are then compared with that of the previous related study by Ameer et al. (2022) to confirm the effectiveness of the proposed method. As seen in Table 17, the method proposed in this study can provide higher accuracy than the previous study that uses the same data to complete a simple Multi-Class Single-Level (MCSL) classification task with limited classes using machine learning, deep learning, and transfer learning approaches. The study by Ameer et al. (2020) employed four machine learning classifiers (RF, Linear SVM, NB, and LR), five deep learning classifiers (GRU, BiGRU, CNN, LSTM, and BiLSTM), and three pre-trained transfer learning models (BERT, XLNet, and RoBERTa). Furthermore, the study applied the most commonly used pre-processing techniques in classification tasks, i.e., lowercasing the post text, removing punctuation marks, removing stop words, and normalising elongated words. When combined with RoBERTa, these techniques yield the best accuracy of 0.83. By comparing this result with those of this study for the same classifier, namely RoBERTa, it can be seen that the RoBERTa classifier cannot handle noises, such as emojis and emoticons, which are still present in the data.

Table 17

Comparison of Results

Study	Classification	Best accuracy
The proposed method	MCML level 1	0.9899
	MCML level 2	0.85
Ameer et al. (2022)	MCSL	0.83

The performance of the RoBERTa classifier is improved due to the modified pre-processing technique proposed in this study, which takes into account such noises, and the best overall results are obtained in the completion of both MCML level 1 and 2 classification tasks. In other words, the proposed MCML classification algorithm produces better results when combined with the appropriate pre-processing technique. For this purpose, the pre-processing techniques used on the data must consider the characteristics of the data. The success of the classifier in completing the classification task using the MCML classification algorithm with fairly good evaluation results for each level suggests that the MCML classification algorithm can perform detailed classification. Furthermore, the MCML classification algorithm is also proven to obtain better classification performance compared to the MCSL classification, which has limitations in the number of classes and classification levels considered for classification. In the MCML classification algorithm, the task is divided into several subtasks. Thus, the classification of mental illness texts is more specific because the texts that indicate a certain disorder are further analysed at more than one classification level, resulting in complex classification results. By grouping texts into mental illness and non-mental illness categories, the MCML classification algorithm can be implemented as a benchmark for text-based early detection of mental illness. If the text is classified as having indications of mental illness, it can be analysed further to determine the type of mental illness.

CONCLUSION AND FUTURE STUDIES

This study explores an intelligent Multi-Class Multi-Level (MCML) classification algorithm for classifying mental health disorders based on Reddit textual data. The MCML classification algorithm is applied using three approaches: machine learning, deep learning, and transfer learning. Two pre-processing scenarios, namely scenario X and scenario Y, are proposed to handle unstructured text data. The proposed method has proven its ability to perform detailed classification with better classification performance using the MCML classification algorithm. This is demonstrated by the success of the classifiers in completing the classification task with fairly good evaluation results for both MCML level 1 and level 2 classification tasks. The performance of each classifier is influenced by the pre-processing scenario used, which is applied to data by considering the characteristics of the data itself. The scenario Y pre-processing technique is more appropriate than the scenario X pre-processing technique, allowing the MCML classification algorithm to perform better and the RoBERTa classifier to obtain the highest overall results that outperform that of the previous study.

This study proposes an intelligent prediction system to overcome the limitations of previous studies in terms of classification levels and number of classes considered for classification. Future studies regarding the classification of mental illnesses using the MCML classification algorithm based on social media textual data can be extended by considering other classification levels with an emphasis on the determination of the severity of mental illness and the possibility of suicide attempts based on the texts analysed. Additionally, this present study remains challenging in terms of low computational memory as the processed social media textual data are derived from the Reddit platform, which has rather long texts. Therefore, potential future studies can use dimensionality reduction methods to create low-dimensional feature spaces and adapt ensemble modelling to improve classification performance.

ACKNOWLEDGMENT

This work was funded by Institut Teknologi Sepuluh Nopember (ITS) under the Departmental Research Funding (*Penelitian Dana Departemen*) Program and the project scheme of the Publication Writing and Intellectual Property Rights (*Penulisan Publikasi dan Hak Kekayaan Intelektual*/PPHKI) Incentive Program; and the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia under the Higher Education Applied Research Excellence (*Penelitian Terapan Unggulan Perguruan Tinggi*/PTUPT) Program.

REFERENCES

Al Hamoud, A., Hoenig, A., & Roy, K. (2022). Sentence subjectivity analysis of a political and ideological debate dataset using LSTM and BiLSTM with attention and GRU models. *Journal of King Saud University - Computer and Information Sciences*, *34*(10), 7974–7987. https://doi.org/10.1016/j.jksuci.2022.07.014

Ameer, I., Arif, M., Sidorov, G., Gòmez-Adorno, H., & Gelbukh, A. (2022). *Mental illness classification on social media texts using deep learning and transfer learning*. http://arxiv.org/abs/2207.01012

Amurark. (2021). *Reddit Dataset*. Github. https://github.com/amurark/mental-health-classification

Bae, Y. J., Shim, M., & Lee, W. H. (2021). Schizophrenia detection using machine learning approach from social media content. *Sensors*, *21*(17), 5924. https://doi.org/10.3390/s21175924

Casola, S., Lauriola, I., & Lavelli, A. (2022). Pre-trained transformers: An empirical comparison. *Machine Learning with Applications*, *9*, 100334. https://doi.org/10.1016/j.mlwa.2022.100334

Chang, M.-Y., & Tseng, C.-Y. (2020). Detecting social anxiety with online social network data. *2020 21st IEEE International Conference on Mobile Data Management (MDM)*, *2020-June*, 333–336. https://doi.org/10.1109/MDM48529.2020.00073

Coppersmith, G., Harman, C., & Dredze, M. (2014). Measuring post traumatic stress disorder in Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 579–582. https://doi.org/10.1609/icwsm.v8i1.14574

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. http://arxiv.org/abs/1810.04805

Duong, H. T., & Nguyen-Thi, T. A. (2021). A review: Pre-processing techniques and data augmentation for sentiment analysis. *Computational Social Networks*, *8*(1). https://doi.org/10.1186/s40649-020-00080-x

Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preoţiuc-Pietro, D., Asch, D. A., & Schwartz, H. A. (2018). Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, *115*(44), 11203–11208. https://doi.org/10.1073/pnas.1802331115

Gaye, B., Zhang, D., & Wulamu, A. (2021). Sentiment classification for employees reviews using regression vector- stochastic gradient descent classifier (RV-SGDC). *PeerJ Computer Science*, *7*, e712. https://doi.org/10.7717/peerj-cs.712

HaCohen-Kerner, Y., Miller, D., & Yigal, Y. (2020). The influence of pre-processing on text classification using a bag-of-words representation. *PLOS ONE*, *15*(5), e0232525. https://doi.org/10.1371/journal.pone.0232525

Hameed, N., Shabut, A. M., Ghosh, M. K., & Hossain, M. A. (2020). Multi-class multi-level classification algorithm for skin lesions classification using machine learning techniques. *Expert Systems with Applications*, *141*, 112961. https://doi.org/10.1016/j.eswa.2019.112961

Hosseini, P., Khoshsirat, S., Jalayer, M., Das, S., & Zhou, H. (2022). Application of text mining techniques to identify actual wrong-way driving (WWD) crashes in police reports. *International Journal of Transportation Science and Technology*. https://doi.org/10.1016/j.ijtst.2022.12.002

Hsu, B.-M. (2020). Comparison of supervised classification models on textual data. *Mathematics*, *8*(5), 851. https://doi.org/10.3390/math8050851

Iyortsuun, N. K., Kim, S.-H., Jhon, M., Yang, H.-J., & Pant, S. (2023). A review of machine learning and deep learning approaches on mental health diagnosis. *Healthcare*, *11*(3), 285. https://doi.org/10.3390/healthcare11030285

Khattak, F. K., Jeblee, S., Pou-Prom, C., Abdalla, M., Meaney, C., & Rudzicz, F. (2019). A survey of word embeddings for clinical text. In *Journal of Biomedical Informatics: X* (Vol. 4). Academic Press Inc. https://doi.org/10.1016/j.yjbinx.2019.100057

Kim, J., Lee, D., & Park, E. (2021). Machine learning for mental health in social media: Bibliometric study. *Journal of Medical Internet Research*, *23*(3), e24870. https://doi.org/10.2196/24870

Kim, J., Lee, J., Park, E., & Han, J. (2020). A deep learning model for detecting mental illness from user content on social media. *Scientific Reports*, *10*(1), 11846. https://doi.org/10.1038/s41598-020-68764-y

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). *ALBERT: A lite BERT for self-supervised learning of language representations*. http://arxiv.org/abs/1909.11942

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A robustly optimised BERT pre-training approach*. http://arxiv.org/abs/1907.11692

Luo, X. (2021). Efficient English text classification using selected machine learning techniques. *Alexandria Engineering Journal*, *60*(3), 3401–3409. https://doi.org/10.1016/j.aej.2021.02.009

Maryamah, M., Arifin, A. Z., Sarno, R., Indraswari, R., & Sholikah, R. W. (2021). Pseudo-relevance feedback combining statistical and semantic term extraction for searching Arabic documents. *International Journal of Intelligent Engineering and Systems*, *14*(5), 238–246. https://doi.org/10.22266/ijies2021.1031.22

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2017). *Advances in pre-training distributed word representations*. http://arxiv.org/abs/1712.09405

Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, *11*(1), 81. https://doi.org/10.1007/s13278-021-00776-6

Noori, B. (2021). Classification of customer reviews using machine learning algorithms. *Applied Artificial Intelligence*, *35*(8), 567–588. https://doi.org/10.1080/08839514.2021.1922843

Rai, N., Kumar, D., Kaushik, N., Raj, C., & Ali, A. (2022). Fake news classification using transformer based enhanced LSTM and BERT. *International Journal of Cognitive Computing in Engineering*, *3*, 98–105. https://doi.org/10.1016/j.ijcce.2022.03.003

Ramírez-Cifuentes, D., Freire, A., Baeza-Yates, R., Puntí, J., Medina-Bravo, P., Velazquez, D. A., Gonfaus, J. M., & Gonzàlez, J. (2020). Detection of suicidal ideation on social media: Multimodal, relational, and behavioural analysis. *Journal of Medical Internet Research*, *22*(7), e17758. https://doi.org/10.2196/17758

Rehm, J., & Shield, K. D. (2019). Global burden of disease and the impact of mental and addictive disorders. *Current Psychiatry Reports*, *21*(2), 10. https://doi.org/10.1007/s11920-019-0997-0

Ríssola, E. A., Aliannejadi, M., & Crestani, F. (2022). Mental disorders on online social media through the lens of language and behaviour: Analysis and visualisation. *Information Processing & Management*, *59*(3), 102890. https://doi.org/10.1016/j.ipm.2022.102890

Ríssola, E. A., Bahrainian, S. A., & Crestani, F. (2019). Anticipating depression based on online social media behaviour. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 11529 LNAI* (pp. 278–290). Springer Verlag. https://doi.org/10.1007/978-3-030-27629-4_26

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. http://arxiv.org/abs/1910.01108

Sarker, I. H. (2021). Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions. *SN Computer Science*, *2*(6), 420. https://doi.org/10.1007/s42979-021-00815-1

Stein, D. J., Palk, A. C., & Kendler, K. S. (2021). What is a mental disorder? An exemplar-focused approach. *Psychological Medicine*, *51*(6), 894–901. https://doi.org/10.1017/S0033291721001185

Su, C., Xu, Z., Pathak, J., & Wang, F. (2020). Deep learning in mental health outcome research: A scoping review. *Translational Psychiatry*, *10*(1), 116. https://doi.org/10.1038/s41398-020-0780-3

Sun, Y., Li, Y., Zeng, Q., & Bian, Y. (2020). Application research of text classification based on random forest algorithm. *2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)*, 370–374. https://doi.org/10.1109/AEMCSE50948.2020.00086

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

Taha, A. Y., Tiun, S., Abd Rahman, A. H., & Sabah, A. (2021). Multilabel over-sampling and under-sampling with class alignment for imbalanced multilabel text classification. *Journal of Information and Communication Technology*, *20*(3), 423–456. https://doi.org/10.32890/jict2021.20.3.6

Uban, A. S., Chulvi, B., & Rosso, P. (2021). An emotion and cognitive based analysis of mental health disorders from social media data. *Future Generation Computer Systems*, *124*, 480–494. https://doi.org/10.1016/j.future.2021.05.032

Vinod Prakash, & Dharmender Kumar. (2023). A modified gated recurrent unit approach for epileptic electroencephalography classification. *Journal of Information and Communication Technology*, *22*(4), 587–617. https://doi.org/10.32890/jict2023.22.4.3

Zhang, T., Schoene, A. M., Ji, S., & Ananiadou, S. (2022). Natural language processing applied to mental illness detection: A narrative review. *Npj Digital Medicine*, *5*(1), 46. https://doi.org/10.1038/s41746-022-00589-7