ISSN 2007-9737

Mental Illness Classification on Social Media Texts Using Deep Learning and Transfer Learning

Muhammad Arif[1], Iqra Ameer[2], Necva Bölücü[3], Grigori Sidorov[1,*], Alexander Gelbukh[1], Vinnayak Elangovan[3]

[1] Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico

[2] The Pennsylvania State University at Abington, Division of Science and Engineering, USA

[3] Commonwealth Scientific and Industrial Research Organisation, Data61, Australia

{arifmuhmmand, necvaa}@gmail.com, {iqa5148, vue9}@psu.edu, {sidorov, gelbukh}@cic.ipn.mx

Abstract. Given the current social distance restrictions across the world, most individuals now use social media as their major medium of communication. Due to this, millions of people suffering from mental diseases have been isolated, and they are unable to get help in person. They have become more reliant on online venues to express themselves and seek advice on dealing with their mental disorders. According to the World Health Organization (WHO), approximately 450 million people are affected. Mental illnesses, such as depression, anxiety, etc., are immensely common and have affected an individual's physical health. Recently, Artificial Intelligence (AI) methods have been presented to help mental health providers, including psychiatrists and psychologists, in decision-making based on patients' authentic information (e.g., medical records, behavioral data, social media utilization, etc.). AI innovations have demonstrated predominant execution in numerous real-world applications, broadening from computer vision to healthcare. This study analyzed unstructured user data on the Reddit platform and classified five common mental illnesses: depression, anxiety, bipolar disorder, ADHD, and PTSD. In this paper, we proposed a Transformer model with late fusion methods to combine the two texts (title and post) of the dataset into the model to detect the mental disorders of individuals. We compared the proposed models with traditional machine learning, deep learning, and transfer learning multi-class models. Our proposed Transformer model with the late fusion method outperformed (F1 score = 89.65) the state-of-the-art performance (F1 score = 89 [35]). This effort will benefit the public health system by automating the detection process and informing the appropriate authorities about people who need emergency assistance.

Keywords. Mental illnesses classification, transformer, late fusion, machine learning, deep learning, transfer learning, Reddit.

1 Introduction

According to certain studies, mental illness can impair a person's physical health as well as her/his intellect, feelings, and behavior (or all three) [50, 32].

450 million people are affected by mental health problems such as depression, schizophrenia, attention-deficit hyperactivity disorder (ADHD), autism spectrum disorder (ASD), etc. [50]. Early diagnosis of mental illness is a fundamental step in better understanding mental health problems and providing care.

Mental illness is usually diagnosed based on self-reporting by individuals in specific surveys

Table 1. Sample instances of Reddit corpus

No.	Reddit Post	Label
1	all the ideas that normally disappear as soon as we reach for a writing device will be captured and started. imagine all the projects we will begin and never finish!	ADHD
2	i know this is long and i don't know if a lot of people will read this but i really just want to help. i had 2 panic attacks over the end of february and first day of march. i went to the doctor and had my blood work	Anxiety
3	for example, did you ever notice that you had manic, hypomanic, depressive, etc. episodes? did you ever notice that sometimes you were ßad" and other times you were ëxcessively happy"? i'm in a sticky	Bipolar
4	i just feel so trapped and i *have* to do something about it. i don't know where i'll go or what i'll do to get by. i just can't stay here any longer.	Depression
5	this is probably going to incite a lot of disagreement, maybe even anger, but that's okay; i'm going to say it anyway. anyone else tired of being told that just talking about your problems will solve your ptsd?	PTSD
6	synesthesia. what is synesthesia? according to google, ßynesthesia is a condition in which one sense (for example, hearing) is simultaneously perceived as if by one or more additional senses such as sight.	None

designed to diagnose specific patterns of feelings or social interactions, in contrast to the diagnosis of other chronic illnesses, which are based on tests and measurements in research settings [19].

In these uncertain times, with COVID-19 torments the world, many people have indicated clinical anxiety or depression. This could be due to lockdown, limited social activities, higher unemployment rates, economic depression, and work-related fatigue.

American Foundation for Suicide Anticipation reported that individuals encounter anxiety (53%) and sadness (51%) more regularly now than before COVID-19 was widespread. Within the past decade, social media has changed social interaction.

In addition to sharing data and news, people share their daily activities, experiences, hopes, emotions, etc., generating reams of data online.

This textual data provides information that can be utilized to design systems to predict people's mental health. Moreover, the current limited social interaction state has forced people to express their thoughts on social media.

In addition, because social interaction is currently limited, people are compelled to express their thoughts on social media. It gives people an open stage to share their opinions with others to find help [35].

Studies that address mental illness primarily utilized deep learning [42, 35] and traditional machine learning [36, 8] models. Recently, Transformer models [47] have gained attention with improvements in Natural Language Processing (NLP) [11, 16, 49] and Computer Vision (CV) [51, 24].

In this work, we adopted the Transformer model that is encoder part of the vanilla Transformer [47] to encode multi text (title and post) simultaneously. We theorize that encoding multiple texts with the same model can improve the quality of the mental illness problem.

We also conducted extensive experiments on late fusion methods to merge the outputs of the proposed model efficiently. We also applied traditional Machine Learning (ML), Deep Learning (DL), and Transfer Learning (TL) approaches to compare the proposed model for automatically detecting mental disorders in social media texts.

Used reddit.com[1] user data proposed by Murarka and Radhakrishnan [35] to determine

[1] https://www.reddit.com/ Last visited: 25-09-2023

Table 2. Number of posts for each mental illness classes in the train/dev/test datasets

Class	Train	Dev	Test
ADHA	2,465	248	248
Anxiety	2,422	248	248
Bipolar	2,407	248	248
Depression	2,450	248	248
PTSD	2,001	248	248
None	1,982	248	248
Total	13,727	1,488	1,488

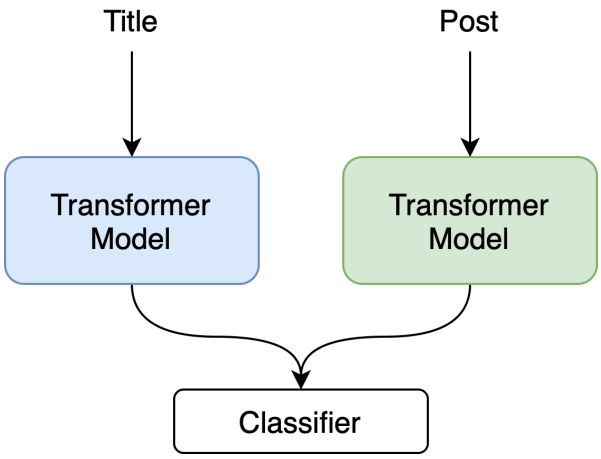

Fig. 1. Late fusion for the mental illness problem

mental illness, Table 1 represents instances of the dataset. The rest of the paper is structured as follows: Section 2 describes the studies on mental illness in literature. Sectio 3 explains the problem and gives dataset insights.

Section 4 gives details of the methodology applied to detect mental disorders with baseline models. Section 5 presents results and their analysis. Section 6 concludes the paper with possible future work.

2 Related Work

Recently, individuals have been using social media to communicate and seek advice on mental health issues. This has motivated researchers to take the information and apply various NLP and ML approaches to help individuals who may want assistance. Initially, many researchers have focused on Twitter text [37, 7, 10], later on the focus has shifted on Reddit platform [25, 17, 7, 52].

A wide range of approaches has been applied to mental health text analysis, from traditional ML to advanced DP. ML points to creating computational algorithms or statistical models capable of extracting hidden patterns from data [39, 44].

For a long time, an increasing number of ML models have been created to analyze healthcare data [36, 8]. Traditional ML approaches require a significant amount of feature engineering for ideal performance, an essential step for most application scenarios to obtain excellent performance and time [15].

Contextual content is created using words. Important insights into text classification can be gained from its structure and order [6, 2]. In the literature, several researchers have extracted the word n-grams to classify user content in social media. [25] used the word n-grams to detect mental illness from Reddit posts.

Another study [23] utilized word n-grams to generate and evaluate artificial mental health records for NLP. According to Coppersmith et al. [10], they employed character-level language models to see how probable a user with mental health concerns would create a series of characters.

Benton et al. [7] determined different types of mental health disorders by applying neural MTL, regression, and multi-layer perceptron single-task learning (STL) models.

Abussa et al. [1] trained the Support Vector Machines to distinguish 200 text messages into two classes: “ADHD or not.” The most crucial step was eliminating the acronym ADHD from the messages before learning, and further information concerning attention disorders was removed from the texts.

The goal was to see how well the Support Vector Machine learns when keywords and even semantically relevant material are unavailable. Deep feed-forward neural network has outperformed typical ML models in a variety of data mining tasks [5, 3, 2, 4], and it has been used

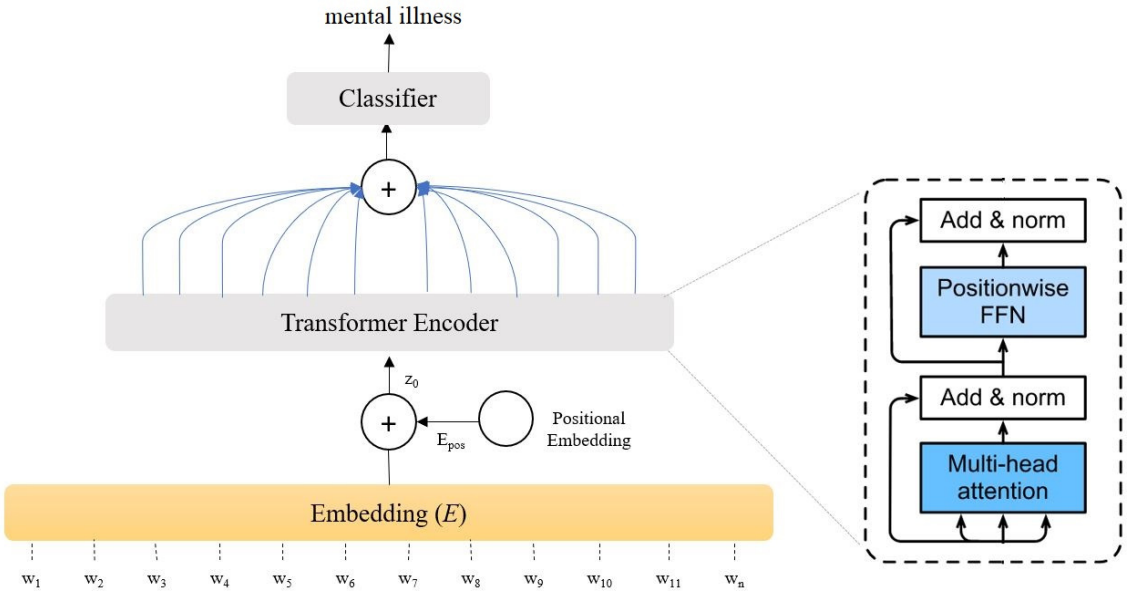

Fig. 2. Overview of the transformer architecture for mental illness problem

in the study of clinical and genetic data to predict mental health disorders.

To diagnose depression, Orabi et al. [37] used word embeddings in combination with a range of neural network models such as CNNs and RNNs. To conduct binary classification on mental health textual posts, Gkotsis et al. [17] used Feed Forward Neural Networks, CNNs, traditional ML such as Support Vector Machine, and Linear classifiers. Sekulic and Strube [41] detected depression, ADHD, anxiety, and other types of mental illnesses by training a binary classifier for each disease with Hierarchical Attention Networks.

The most recent work on this was a CNN-based classification model Kim et al. [25]. The team trained a separate binary classifier for each type of mental disorder to conduct the detection. Hu and Sokolova [21] found the potential factors to influence a person's mental health during the Covid-19 pandemic by applying ML classifiers such as Naive Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and Gradient Boosting (GB).

They have also presented an analysis of the feature selection technique called LIME (Local, Interpretable Model-agnostic Explanations) [40]. In a recent study, Shatte et al. [42] applied ML techniques to the mental health domain.

They reviewed the literature using four key application domains—detection and diagnosis, prognosis, treatment and support, public health applications, and research and clinical administration. Another research examined the recently developed field of DL methods in psychiatry. They concentrated on DL and integrated statistical ML correlations with semantically interpretable computer models of brain dynamics or behaviour [14].

A variety of cutting-edge NN models were employed in DL-based methods. Shared task CLPsych[2] series played a significant part in developing mental health detection. CNN, RNN, LSTM, and BiLSTM were found to be the most commonly applied models.

In today's research world, TL is extremely important. Researchers attempt to acquire greater accuracy and performance in several research studies by using several types of transformers. Murarka et al. [35] examined three approaches for identifying and diagnosing mental illness

[2]clpsych.org/ Last visited: 25-09-2023

Table 3. Parameter settings of models

HyperParameter	CNN	LSTM	BiLSTM	Transformer
learning rate	$1e-5$	$1e-5$	$1e-5$	$1e-5$
batch size	16	16	16	16
# of LSTM layers	-	1	1	-
hidden units	-	250	250	250
dropout	0.1	0.2	0.2	0.2
# of Kernel	2 (3,4)	-	-	-
# of layers	-	-	-	2
# of hidden	-	-	-	2
d model	-	-	-	2
dropout	-	-	-	0.2
heads	-	-	-	2

Table 4. Results with late fusion methods

Method	Precision	Recall	F1 score
Concatenation	**89.86**	**89.58**	**89.65**
Average	88.06	87.70	87.78
Weighted Average	86.68	85.69	85.85
Maximum	87.81	87.50	87.60
Minimum	88.04	87.84	87.86

on the Reddit dataset, including LSTM, BERT, and RoBERTa.

RoBERTa outperformed the other two methods. Dhanalaxmi et al. [12] employed RoBERTa to categorize COVID-19-related informative tweets, and their method yielded the best results. Mathur et al. [33] applied LSTM with an attention mechanism to estimate suicidal intent using temporal psycholinguistics.

Shickel et al. [43] utilized a deep transfer learning model to predict emotional valence in mental health text and achieved the highest performance with BERT. Moreover, they claimed that in automatic mental health systems, where labeled data is frequently scarce, recent transfer learning algorithms should become a crucial component.

Du et al. [13] looked at approaches for identifying suicide-related psychiatric stresses in Twitter data using deep learning-based approaches and a transfer learning approach that uses an existing clinical text annotation dataset. They demonstrated the advantages of deep learning-based techniques compared to conventional machine learning algorithms.

Additionally, it was discovered that the transfer learning technique might potentially reduce annotation work and further improve performance. To automatically detect public opinions, behavioral intentions, and attitudes concerning COVID-19 vaccinations from Tweets, Cagliero and Garza [29] used transfer learning with a pre-trained BERT model.

They showed that transfer learning models outperformed traditional machine learning models. To summarize, ML and DL techniques have been used in health care problems as efficient methods using text on social media platforms due to their ability to outperform naive learning models significantly [9].

Motivated by these models, we proposed a Transformer model with late fusion methods to combine the title and post of the dataset into the model to detect the mental disorders of individuals. To the best of our knowledge, none of the prior studies have applied the Transformer model with late fusion models for the mental illness problem.

3 Problem Description and Dataset

3.1 Mental Illness Problem

Mental illness problem is a multi-class classification problem where a given text is classified into one of the six following mental disorder classes:

– **ADHD**: A mental condition that impairs your ability to focus, maintain stillness, and control your actions (common in children)[3].

– **Anxiety**: A feeling of uneasiness, fear, and dread[4].

[3] www.cdc.gov/ncbddd/adhd/facts.html Last visited: 125-09-2023

[4] /medlineplus.gov/anxiety.html\#:~:text=Anxiety\%20is\%20a\%20feeling\%20of,before\%20making\%20an\%20important\%20decision Last visited: 25-09-2023

Table 5. Confusion matrix of the transformer model with concatenation late fusion method

		Predicted					
		ADHD	**Anxiety**	**Bipolar**	**Depression**	**PTSD**	**None**
True	**ADHD**	224	10	5	6	3	0
	Anxiety	1	222	1	19	5	0
	Bipolar	9	8	211	16	3	1
	Depression	3	8	12	219	6	0
	PTSD	0	19	9	7	213	0
	None	0	1	1	1	1	244

– **Bipolar**: Extreme mood swings, including emotional highs and lows, are a symptom of a mental health issue [5].

– **Depression**: A widespread and significant medical condition that has a negative impact on how someone feels, thinks, and acts[6].

– **PTSD**: A condition that some people experience after going through a stressful, terrifying, or deadly experience[7].

– **None**: No mental illness.

3.2 Dataset

Murarka et al. [35] developed a benchmark multi-class dataset from the Reddit social media platform for mental illness detection.

The dataset comprises a total of $16,703$ posts. The dataset was further divided into training, development, and test sets.

Table 2 presents the number of posts for each mental illness class.

[5] www.mayoclinic.org/diseases-conditions/bipolar-disorder/symptoms-causes/syc-20355955 Last visited: 25-09-2023.

[6] www.psychiatry.org/patients-families/depression/what-is-depression Last visited: 25-09-2023.

[7] www.nimh.nih.gov/health/topics/post-traumatic-stress-disorder\\-ptsd\#:~:text=Post\%2Dtraumatic\%20stress\%20disorder\%20(PTSD,danger\%20or\%20to\%20avoid\%20it. Last visited: 25-09-2023.

4 Model

4.1 Transformer Model

Transformer model [47] is gaining interest due to state-of-the-art performance in NLP tasks such as machine translation [48, 30], and sequence tagging [46, 20]. The Transformer model comprises encoder-decoder architectures that process sequential data in parallel without a recurrent network.

Instead of paying attention to the last state of the encoder, as is common with RNNs, the encoder architecture in Transformer extracts information from the whole sequence. This allows the decoder to assign greater weight to a certain input element for each output element.

In this study, we proposed Transformer models based on the vanilla Transformer proposed by Vaswani et al. [47] and used the encoder module of the Transformer to perform classification by mapping the data to the mental illness classes. The architecture of the Transformer model is shown in Figure 2.

Let $S = \{X_i, W_i, m_i\}_{i=1}^{T}$ denote a set of T samples, where X_i is a title, W_i is a post. m_i is the corresponding mental illness class (adhd, anxiety, bipolar, depression, ptsd, none). The words $\{w_1, w_2, \cdots, w_n\}$ for a text, which can be *title* or *post*, are mapped to the corresponding embeddings in the embedding layer, and the positional information E_{pos} is encoded and appended to the text representation and fed into the encoder layer, which consists L identical layers. The classification layer is a

Table 6. Results of the proposed model with baseline models

Model	**F1**	**Precision**	**Recall**
Transformer	**89.65**	**89.86**	**89.58**
Classical Machine Learning			
ML Algorithm	**F1**	**Precision**	**Recall**
LinearSVC	77.18	77.66	77.15
LR	77.87	78.24	77.89
NB	66.49	72.18	66.73
RF	70.85	72.46	70.50
Deep Learning			
DL Algorithm	**F1**	**Precision**	**Recall**
CNN	81.64	82.84	82.65
LSTM	83.73	84.10	83.60
BiLSTM	83.84	84.06	83.74
Transfer Learning			
TL Algorithm	**F1**	**Precision**	**Recall**
BERT	80.82	80.87	80.85
AlBERT	80.45	80.90	80.38
RoBERTa	84.41	85.10	84.41
State-of-the-Art			
Method	**F1**	**Precision**	**Recall**
RoBERTa	89	89	89

softmax layer that takes the average of the last transformer encoder layer o and multiplies the corresponding weights to get classification:

$$\hat{s} = \text{softmax}(W \times o + b), \tag{1}$$

where $\hat{s}$ is the predicted result through the model, W is the weighted matrix, and b is the bias.

How do title and post contribute to the predictions? Over the years, various fusion techniques (e.g., early fusion or late fusion) have been developed for prediction in computer vision [22, 18] and NLP tasks [45, 34].

Since there are two parts for each instance (title and post), we also applied late fusion combining the outputs of each model at the classification layer. Moreover, we tried various combinations of methods in experimental settings (e.g., concatenation, average, maximum, minimum, weighted average).

4.2 Baseline Models

Since the mental illness dataset used in this study is relatively new, we applied ML, DL, and TL algorithms to get baseline scores. The models are summarised as follows:

- **Machine Learning Classifiers:** We applied four different ML classifiers, including Random Forest, Linear Support Vector Machine, Multinomial Naive Bayes, and Logistic Regression using sckit-learn library[8].
- **Deep Learning Methods:** We applied base DL models: LSTM, BiLSTM, and CNN. The pre-trained embeddings were used as the input layer, and the softmax layer as the output layer of the models.
- **Transfer Learning Methods:** Transformer-based pre-trained language models (PLMs) such as BERT [11], RoBERTa [31], AlBERT [27] have shown state-of-art performance in many down-stream NLP tasks. The PLMs used in NLP problems, called transfer learning models, yielded top results in various NLP tasks without critical task-specific design changes [28, 11]. We employed the BERT, AlBERT, and RoBERTa models in this study.

5 Results and Analysis

5.1 Experimental Setting

We implemented the proposed DL and TL models using the PyTorch library [38]. The Adam optimizer [26] was used with an epsilon value of $1e-8$ and the default max grad norm. We used early stopping with 5 patience.

We utilized pre-trained language models (BERT [11], RoBERTa [31], etc.) to convert words into embeddings. To tokenize the words, we set the maximum length 35 and 512 for the title and post, respectively, for all pre-trained language models (BERT, RoBERTa, etc.).

[8] scikit-learn.org/ Last visited: 125-09-2023.

Table 7. Results of the proposed model with baseline models

	Title			Post		
Model	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**
Transformer	70.09	70.46	70.09	83.63	83.90	83.53
ML Algorithms	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**
Classical Machine Learning						
LinearSVC	65.48	65.78	65.52	77.18	77.66	77.15
LR	65.37	66	65.52	77.87	78.24	77.89
NB	62.46	68.56	62.37	66.49	72.18	66.73
RF	61.63	62.02	61.63	70.85	72.46	70.50
Deep Learning						
DL Algorithm	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**
CNN	69.94	70.85	69.76	81.64	82.84	82.65
LSTM	71.46	72.07	71.44	83.73	84.10	83.60
BiLSTM	70.12	70.65	69.89	83.84	84.06	83.74
Transfer Learning						
DL Algorithm	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**
BERT	70.06	70.29	69.96	80.82	80.87	80.85
AlBERT	67.37	67.58	67.34	80.45	80.90	80.38
RoBERTa	70.68	71.27	70.56	**84.41**	**85.10**	**84.41**

For TL models, we added an output layer with a softmax function for training and set the learning rate to $1e-5$ and the batch size to 16.

In ML models, the number of features in each experiment was set to 1,000, i.e., we used the n-grams with the highest TF-IDF values. For the combination of word n-grams, the length of N was minimum = 1 and maximum = 3.

Since the dataset was already pre-processed by eliminating URLs or usernames containing sensitive material, we did not apply any pre-processing techniques before classification. We fine-tuned the models using the development set of the dataset.

Table 3 shows the parameter settings of DL and the proposed transformer models. We evaluated the models using the following three metrics: micro precision, micro recall, and micro F1-Score.

5.2 Main Results

Table 6 presents the proposed models' results and comparison with baseline models. In this Table, "ML Algorithms" indicates traditional ML algorithms. The "LinearSVC" indicates Linear Support Vector Classifier, "LR" indicates Logistic Regression, "NB" indicates Naive Bayes, and "RF" indicates Random Forest classifier.

The "DL Algorithms" indicates DL algorithms used in this study, such as CNN, LSTM, and BiLSTM. The "TL Algorithms" refer to pre-trained TL algorithms applied to evaluate Reddit corpus, i.e., BERT, XLNet, AlBERT, and RoBERTa. Using traditional ML algorithms, overall, best results (F1 = 77.87) are obtained using a combination of word n-grams when the length of *N* was minimum = 1, maximum = 3 with the Logistic Regression model. This shows that combinations of word grams (length of *N* = 1-3) were the most suitable

Table 8. Class-wise results

	Title			Post			Title + Post		
Class	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
ADHD	68	77	72	85	87	86	95	90	92
Anxiety	60	69	64	72	83	77	83	90	86
Bipolar	64	55	59	82	78	80	88	85	87
Depression	69	65	67	80	77	78	82	88	85
PTSD	67	65	66	86	83	84	82	88	85
None	95	90	92	98	94	96	82	88	85

features when we trained the model on the Reddit social media platform.

In DL models, the overall best results are achieved with BiLSTM (F1 = 83.84), which shows that DL models are suitable for detecting mental illness. Additionally, DL results are almost similar to the results of TL models.

Since RoBERTa pre-trained model in TL methods yielded the best results, we used RoBERTa pre-trained embeddings as the input layer of the DL models (CNN, LSTM, and BiLSTM) and the proposed Transformer model.

The state-of-the-art RoBERTa [35] model was trained on title + post text, which is different from our RoBERTa model as we trained it on posts only. Among the baseline models (ML, DL, and TL), RoBERTa outperformed the traditional ML and DL models with an F1 score of 84.41 on this challenging multi-class mental illness detection problem.

Overall, we obtained the highest score with the proposed Transformer model with the `concatenation` late fusion method (F1 = 89.65). Our proposed model outperformed the state-of-the-art RoBERTa [35] model (F1 = 89).

Table 5 shows the confusion matrix of the proposed Transformer model with the concatenation late fusion method. The model is good at predicting non-illness samples. However, it confuses at prediction of the classes anxiety, bipolar, depression, and ptsd.

The terms `Depression` and `Anxiety` are presented in data instances of the `ADHD` and `PTSD` classes more than these class themselves. One could expect poor outcomes due to this, but these classes outperformed all others.

This exhibits the actual potential of our approach since it does not depend solely on the mention of class names in the post but also has a deep awareness of the post's context.

5.3 Data

To understand the impact of the dataset comprising titles and posts, we performed experiments with the proposed Transformer and the baseline models using title and post separately. Table 7 represented the F1, Precision, and Recall scores of Transformer, traditional ML, and DL models.

We obtained the best results with the posts (F1 = 84.41). We analyzed that the length of the titles in the dataset is shorter than the posts, which indicates that they are not informative enough. Therefore, we can say that the length of the text is important for the models, especially for the DL.

To understand the impact of the methods, Table 8 presented the class-wise results of the Transformer using the title and post separately and concatenating them.

We performed experiments using the proposed Transformer model to get insights on the class-wise performance of our proposed Transformer model on titles and post text separately and by combining them. The Transformer model obtained a 0.96 F1 score for `none` class on posts only, which shows that this proposed model will suffer from minimum false positives in detecting mental illness on social media text. The Transformer model using title and post together increased the F1 score of each

Table 9. Class-wise results of late fusion methods

		ADHD	Anxiety	Bipolar	Depression	PTSD	None
Concatenation	Precision	95	83	88	82	92	100
	Recall	90	90	85	88	86	98
	F1	92	86	87	85	89	99
Average	Precision	94	82	86	78	92	98
	Recall	90	87	83	86	82	97
	F1	92	85	84	82	87	97
Weighted Average	Precision	94	69	88	83	88	98
	Recall	84	91	82	77	84	97
	F1	89	79	85	80	86	97
Maximum	Precision	94	81	81	80	90	100
	Recall	89	86	86	83	84	97
	F1	92	84	84	82	87	98
Minimum	Precision	92	81	87	81	89	98
	Recall	92	86	81	88	82	98
	F1	92	84	84	84	86	98

mental illness class. However, the performance of `none` class decreased compared to other models (Transformer for title and post separately).

The best performing class among the mental disorders was `adhd`, while the performance of the other classes was similar. This shows that the model significantly fits the dataset. The training dataset contains fewer samples of `ptsd` class, and despite this, the F1 score of `ptsd` class was not dropped. The class-wise results were very similar to the RoBERTa [35] model except for the `none` class. Their model performed well on `none` class.

5.4 Late Fusion

To extend the impact of the data on the problem, we applied late fusion (Figure 1). The results of the methods used in late fusion are shown in Table 4. We used RoBERTa [31] pre-trained embeddings in the models with the same parameters for each model (See Table 3 for hyperparameter settings of the models).

The results showed that all methods improved the results compared to the Transformer model using only one input (title or post). We achieved the highest score (F1 = 89.65) by *concatenation* fusion methods. Table 9 presents the class-wise results of the Transformer model with late fusion methods.

It can be observed that the late fusion method of concatenation performed better on all classes than other methods. Moreover, there is not much difference in the performances of the late fusion methods. It can be inferred that the method can be used for datasets containing two or more texts to increase performance.

6 Conclusion

The present Covid-19 outbreak and globally forced isolation were our primary motivations for multi-class mental illness detection efforts. We believe that social media platforms have become the most widely used communication medium for individuals, allowing them to express themselves without fear of judgment.

We applied the Transformer model with fusion methods and state-of-the-art traditional ML, DL, and TL-based methods for multi-class mental illness detection problem. The best results (see

Table 4) were obtained with the Transformer model with concatenation late fusion method (F1 score = 89.65).

In the future, we plan to develop a multi-label mental illness dataset, which would be more reflective of the situation than a multi-class dataset, as a post can have more than one mental disease instead of one per post, i.e., depression and anxiety.

We can also use the data augmentation technique on top of existing mental health data [35]. Moreover, we plan to apply other TL-based models, such as DistilBERT, in the future. An ensemble modeling would also be considered to improve classification performance.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20241816, 20241819, and 20240951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico, and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

1. **Abusaa, M., Diederich, J., Al-Ajmi, A. (2004).** Machine learning, text classification and mental health. 12th National Health Informatics Conference, pp. 1–7.
2. **Ameer, I., Ashraf, N., Sidorov, G., Gómez-Adorno, H. (2020).** Multi-label emotion classification using content-based features in twitter. Computación y Sistemas, Vol. 24, No. 3, pp. 1159–1164.
3. **Ameer, I., Siddiqui, M. H. F., Sidorov, G., Gelbukh, A. (2019).** CIC at SemEval-2019 task 5: Simple yet efficient approach to hate speech detection, aggressive behavior detection, and target classification in Twitter. Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 382–386. DOI: 10.18653/v1/S19-2067.
4. **Ameer, I., Sidorov, G. (2021).** Author profiling using texts in social networks. In Handbook of Research on Natural Language Processing and Smart Service Systems. IGI Global, pp. 245–265. DOI: 10.4018/978-1-7998-4730-4.ch011.
5. **Ameer, I., Sidorov, G., Gomez-Adorno, H., Nawab, R. M. A. (2022).** Multi-label emotion classification on code-mixed text: Data and methods. IEEE Access, Vol. 10, pp. 8779–8789. DOI: 10.1109/ACCESS.2022.3143819.
6. **Ameer, I., Sidorov, G., Nawab, R. M. A. (2019).** Author profiling for age and gender using combinations of features of various types. Journal of Intelligent & Fuzzy Systems, Vol. 36, No. 5, pp. 4833–4843. DOI: 10.3233/JIFS-179031.
7. **Benton, A., Mitchell, M., Hovy, D. (2017).** Multi-task learning for mental health using social media text. CoRR. DOI: 10.48550/arXiv.1712.03538.
8. **Biship, C. M. (2007).** Pattern recognition and machine learning. Springer New York, NY.
9. **Cagliero, L., Garza, P. (2013).** Improving classification models with taxonomy information. Data & Knowledge Engineering, Vol. 86, pp. 85–101. DOI: 10.1016/j.datak.2013.01.005.
10. **Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., Mitchell, M. (2015).** CLPsych 2015 shared task: Depression and PTSD on Twitter. Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pp. 31–39.

11. **Devlin, J., Chang, M., Lee, K., Toutanova, K. (2018).** BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2019, pp. 4171–4186. DOI: 10.48550/arXiv.1810.04805.
12. **Dhanalaxmi, S., Agarwal, R., Sinha, A. (2020).** Detection of COVID-19 informative tweets using RoBERTa. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text, pp. 409–413. DOI: 10.48550/arXiv.2010.11238.
13. **Du, J., Zhang, Y., Luo, J., Jia, Y., Wei, Q., Tao, C., Xu, H. (2018).** Extracting psychiatric stressors for suicide from social media using deep learning. BMC Medical Informatics and Decision Making, Vol. 18, No. 43, pp. 77–87. DOI: 10.1186/s12911-018-0632-8.
14. **Durstewitz, D., Koppe, G., Meyer-Lindenberg, A. (2019).** Deep neural networks in psychiatry. Molecular Psychiatry, Vol. 24, No. 11, pp. 1583–1598. DOI: 10.1038/s41380-019-0365-9.
15. **Dwyer, D. B., Falkai, P., Koutsouleris, N. (2018).** Machine learning approaches for clinical psychology and psychiatry. Annual review of clinical psychology, Vol. 14, pp. 91–118. DOI: 10.1146/annurev-clinpsy-032816-045037.
16. **Gillioz, A., Casas, J., Mugellini, E., Abou-Khaled, O. (2020).** Overview of the transformer-based models for NLP tasks. 2020 15th Conference on Computer Science and Information Systems (FedCSIS), pp. 179–183. DOI: 10.15439/2020F20.
17. **Gkotsis, G., Oellrich, A., Velupillai, S., Liakata, M., Hubbard, T. J., Dobson, R. J., Dutta, R. (2017).** Characterisation of mental health conditions in social media using informed deep learning. Scientific Reports, Vol. 7, No. 1, pp. 1–11. DOI: 10.1038/srep45141.
18. **Gunes, H., Piccardi, M. (2005).** Affect recognition from face and body: Early fusion vs. late fusion. 2005 IEEE international conference on systems, man and cybernetics, IEEE, Vol. 4, pp. 3437–3443. DOI: 10.1109/ICSMC.2005.1571679.
19. **Hamilton, M. (1967).** Development of a rating scale for primary depressive illness. British journal of social and clinical psychology, Vol. 6, No. 4, pp. 278–296. DOI: 10.1111/j.2044-8260.1967.tb00530.x.
20. **He, Z., Wang, Z., Wei, W., Feng, S., Mao, X., Jiang, S. A. (2020).** A survey on recent advances in sequence labeling from deep learning models. ArXiv, Vol. abs/2011.06727. DOI: 10.48550/arXiv.2011.06727.
21. **Hu, Y., Sokolova, M. (2021).** Explainable multi-class classification of the CAMH COVID-19 mental health data. ArXiv. DOI: 10.48550/arXiv.2105.13430.
22. **Ionescu, B., Benois-Pineau, J., Piatrik, T., Quénot, G. (2014).** Fusion in computer vision: Understanding complex visual content. Springer. DOI: 10.1007/978-3-319-05696-8.
23. **Ive, J., Viani, N., Kam, J., Yin, L., Verma, S., Puntis, S., Cardinal, R. N., Roberts, A., Stewart, R., Velupillai, S. (2020).** Generation and evaluation of artificial mental health records for natural language processing. NPJ Digital Medicine, Vol. 3, No. 1, pp. 1–9. DOI: 10.1038/s41746-020-0267-x.
24. **Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., Shah, M. (2021).** Transformers in vision: A survey. ACM Computing Surveys (CSUR), Vol. 54, No. 10s, pp. 1–41. DOI: 10.1145/3505244.
25. **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, Vol. 10, No. 1, pp. 11846. DOI: 10.1038/s41598-020-68764-y.
26. **Kingma, D. P., Ba, J. (2014).** Adam: A method for stochastic optimization. International Conference on Learning Representations. DOI: 10.48550/ARXIV.1412.6980.

27. **Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R. (2019).** Albert: A lite bert for self-supervised learning of language representations. International Conference on Learning Representations.

28. **Li, X., Fu, X., Xu, G., Yang, Y., Wang, J., Jin, L., Liu, Q., Xiang, T. (2020).** Enhancing BERT representation with context-aware embedding for aspect-based sentiment analysis. IEEE Access, Vol. 8, pp. 46868–46876. DOI: 10.1109/ACCESS.2020.2978511.

29. **Liu, S., Li, J., Liu, J. (2021).** Leveraging transfer learning to analyze opinions, attitudes, and behavioral intentions toward COVID-19 vaccines: Social media content and temporal analysis. Journal of Medical Internet Research, Vol. 23, No. 8, pp. e30251. DOI: 10.2196/30251.

30. **Liu, X., Duh, K., Liu, L., Gao, J. (2020).** Very deep transformers for neural machine translation. ArXiv. DOI: 0.48550/arXiv.2008.07772.

31. **Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019).** Roberta: A robustly optimized BERT pretraining approach. The International Conference on Learning Representations 2020 Conference Blind Submission. DOI: 10.48550/arXiv.1907.11692.

32. **Marcus, M., Yasamy, M. T., van-Ommeren, M., Chisholm, D., Saxena, S. (2012).** Depression: A global public health concern. DOI: 10.1037/e517532013-004.

33. **Mathur, P., Sawhney, R., Chopra, S., Leekha, M., Shah, R. R. (2020).** Utilizing temporal psycholinguistic cues for suicidal intent estimation. Advances in Information Retrieval, Springer International Publishing, Vol. 12036, pp. 265–271. DOI: 10.1007/978-3-030-45442-5_33.

34. **Mukherjee, S. (2019).** Deep learning technique for sentiment analysis of hindi-english code-mixed text using late fusion of character and word features. 2019 IEEE 16th India Council International Conference (INDICON), pp. 1–4. DOI: 10.1109/INDICON47234.2019.9028928.

35. **Murarka, A., Radhakrishnan, B., Ravichandran, S. (2021).** Classification of mental illnesses on social media using RoBERTa. Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis, pp. 59–68.

36. **Murphy, K. P. (2012).** Machine learning: A probabilistic perspective. MIT press, Cambridge, Massachusetts, USA.

37. **Orabi, A. H., Buddhitha, P., Orabi, M. H., Inkpen, D. (2018).** Deep learning for depression detection of twitter users. Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, pp. 88–97. DOI: 10.18653/v1/W18-0609.

38. **Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S. (2019).** Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, Vol. 32.

39. **Pervaz, I., Ameer, I., Sittar, A., Nawab, R. M. A. (2015).** Identification of author personality traits using stylistic features: Notebook for pan at clef 2015. CLEF (Working Notes), pp. 1–7.

40. **Ribeiro, M. T., Singh, S., Guestrin, C. (2016).** " Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135–1144. DOI: 10.1145/2939672.2939778.

41. **Sekulic, I., Strube, M. (2020).** Adapting deep learning methods for mental health prediction on social media. Proceedings of the 5th Workshop on Noisy User-generated Text

(W-NUT 2019), pp. 322–327. DOI: 10.18653/v1/D19-5542.

42. **Shatte, A. B., Hutchinson, D. M., Teague, S. J. (2019).** Machine learning in mental health: a scoping review of methods and applications. Psychological medicine, Vol. 49, No. 9, pp. 1426–1448. DOI: 10.1017/S0033291719000151.

43. **Shickel, B., Heesacker, M., Benton, S., Rashidi, P. (2020).** Automated emotional valence prediction in mental health text via deep transfer learning. 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), pp. 269–274. DOI: 10.1109/BIBE50027.2020.00051.

44. **Sittar, A., Ameer, I. (2018).** Multi-lingual author profiling using stylistic features. FIRE (Working Notes), pp. 240–246.

45. **Soriano-Morales, E. P., Ah-Pine, J., Loudcher, S. (2017).** Fusion techniques for named entity recognition and word sense induction and disambiguation. Discovery Science: 20th International Conference, Springer International Publishing, pp. 340–355.

46. **Tsai, H., Riesa, J., Johnson, M., Arivazhagan, N., Li, X., Archer, A. (2019).** Small and practical BERT models for sequence labeling. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3632–3636. DOI: 10.18653/v1/D19-1374.

47. **Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. (2017).** Attention is all you need. Advances in Neural Information Processing Systems, Vol. 30, pp. 5998–6008.

48. **Wang, Q., Li, B., Xiao, T., Zhu, J., Li, C., Wong, D. F., Chao, L. S. (2019).** Learning deep transformer models for machine translation. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1810–1822. DOI: 10.18653/v1/P19-1176.

49. **Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von-Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le-Scao, T., Gugger, S., Drame, M., et al. (2020).** Transformers: State-of-the-art natural language processing. Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6.

50. **World Health Organization (2001).** The world health report 2001: Mental health: New understanding, new hope. https://iris.who.int/handle/10665/42390.

51. **Zhou, D., Kang, B., Jin, X., Yang, L., Lian, X., Hou, Q., Feng, J. (2021).** DeepViT: Towards deeper vision transformer. arXiv. DOI: 10.48550/arXiv.2103.11886.

52. **Zirikly, A., Resnik, P., Uzuner, O., Hollingshead, K. (2019).** CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, pp. 24–33. DOI: 10.18653/v1/W19-3003.

Article received on 26/02/2024; accepted on 12/04/2024.
**Corresponding author is Grigori Sidorov.*