2020 IEEE 39th International Performance Computing and Communications Conference (IPCCC) | 978-1-7281-9829-3/20/$31.00 ©2020 IEEE | DOI: 10.1109/IPCCC50635.2020.9391501

A Multimodal Feature Fusion-Based Method for Individual Depression Detection on Sina Weibo

1st Yiding Wang
College of Cybersecurity
Sichuan University
Chengdu, China
wangyiding@stu.scu.edu.cn

2nd Zhenyi Wang
College of Cybersecurity
Sichuan University
Chengdu, China
wangzhenyi@stu.scu.edu.cn

3rd Chenghao Li
College of Cybersecurity
Sichuan University
Chengdu, China
2017141531008@stu.scu.edu.cn

4th Yilin Zhang
College of Cybersecurity
Sichuan University
Chengdu, China
2017141482187@stu.scu.edu.cn

5th Haizhou Wang*
College of Cybersecurity
Sichuan University
Chengdu, China
whzh.nc@scu.edu.cn

***Abstract*—Existing studies have shown that various types of information on the online social network (OSN) can help predict the early stage of depression. However, studies using machine learning methods to accomplish depression detection tasks still do not have high classification performance, suggesting that there is much potential for improvement in their feature engineering. In this paper, we first construct a dataset on Sina Weibo (a leading OSN with the largest number of active users in the Chinese community), namely the Weibo User Depression Detection Dataset (WU3D). It includes more than 10,000 depressed users and 20,000 normal users, both of which are manually labeled and rechecked by specialists. Then, we extract text-based word features using the popular pretrained model XLNet and summarize nine statistical features related to user text, social behavior, and pictures. Moreover, we construct a deep neural network classification model, i.e. Multimodal Feature Fusion Network (MFFN), to fuse the above-extracted features from different information sources and further accomplish the classification task. The experimental results show that our approach achieves an F1-Score of 0.9685 on the test dataset, which has a good performance improvement compared to the existing works. In addition, we verify that our multimodal detecting approach is more robust than multimodel ensemble ones. Our work could also provide new research methods for depression detection on other OSN platforms.**

***Index Terms*—Depression detection, online social network, feature engineering, deep learning, multimodal fusion.**

I. Introduction

A. Background

Major depressive disorder (MDD), referred to as depression, is a common mental disease. According to a survey of the World Health Organization (WHO)[1], more than 300 million people worldwide suffer from depression. Despite the current availability of psychotherapy, medical therapy, and other modalities for the treatment of depression, 76%-85% of patients in low- and middle-income countries remain untreated. The inability to make an accurate assessment in the early stage of depression leads to a large number of depressed individuals difficult to get diagnosis and treatment timely [1].

Nowadays, people are more frequently using the Online Social Network (OSN) to express opinions and emotions. It provides researchers with a novel and effective way to detect the mood, communication, activity, and social behavior pattern of individuals [2]. Various information published on the OSN is proved to be able to reflect the user's state of mind, which can help researchers to specifically characterize depressed users [1]–[4]. Sina Weibo (hereinafter referred to as "Weibo") is the most popular OSN in the Chinese community with more than 462 million active daily users in 2019 statistics [5].

As artificial intelligence technologies progressing, machine learning approaches have made great contributions to the detection of depression [6]–[9]. Among them, the Multimodal Feature Fusion (MFF) is an approach that jointly considers information from heterogeneous sources and makes target prediction [10]. This approach integrates data from multiple modalities (text, image, video) to eliminate ambiguity and uncertainty through complementary information, which in turn can lead to more accurate classification results.

B. Challenges

However, existing machine learning-based online depression detection approaches still leave many unresolved issues.

Firstly, many previous studies are not user-oriented modeling [11], [12]. Such results cannot be directly applied to user-level depression detection, or it may lead to an incorrect prediction. For example, the model may predict a single tweet as depressive, but cannot determine whether the person is depressed since normal users also express their depressive emotions on the OSN sometimes.

*This work is supported by the National Natural Science Foundation of China (NSFC) under grant nos. 61802270, 61802271, 81602935, and 81773548.

*Corresponding author: H. Wang (whzh.nc@scu.edu.cn)

[1] https://www.who.int/en/news-room/fact-sheets/detail/depression

Secondly, the size of the dataset used for modeling is insufficient [2], [11], [13]–[15], with only a few hundred to a few thousand data samples being used. As a consequence, the trained model can easily overfit the dataset, thus failed to reach good generalization performance.

Moreover, not enough studies of user depression detection have been proposed on Weibo compare to Twitter and Facebook. To the best of our knowledge, currently, there is no published large Weibo user depression detection dataset available for researchers to use.

C. Contributions

Given the above problems and challenges, we hereby summarize the contributions of our work as below:

- **We build and publish a large-scale labeled dataset - Weibo User Depression Detection Dataset (WU3D).**[2] In WU3D, more than 10,000 depressed users and more than 20,000 normal users are collected, labeled, and checked. Each user sample contains enriched information fields, including the user's nickname, tweets, the posting time, posted pictures, the user's gender, etc.
- **We summarize nine features of depressed users, four of which are the first to be proposed.** All of these features have a positive contribution to the classification task, with significant distributional differences between normal and depressed users.
- **We construct a deep neural network (DNN) classification model for depression detection.** It implements a multimodal learning strategy to simultaneously process multiple inputs, including the text-based word features and the manually extracted features. The experimental results show that our proposed model achieves the highest F1-Score and the best robustness compare to other popular classification models.

The subsequent sections of this paper are organized as follows. In Section II, related work and achievements in the field of machine learning-based online depression detection are introduced. The proposed framework is elaborated in Section III. Furthermore, Section IV gives experiments of our proposed detecting model and several classification models that are popular or most commonly used in related studies. At the end of the paper, Section V summarizes this work and discusses directions for future work.

II. Related Work

The work of machine learning-based depression detection on the OSN can be divided into two directions. **(i)** Manually extract features and build Traditional Machine Learning (TML) models for classification. **(ii)** Use Deep Learning (DL) approaches to automatically extract features and construct DNN models as classifiers.

Particularly, some of the detecting approaches based on DL introduces TML methods to further improve their model performance. The work of each research direction will be described respectively below.

A. Traditional Machine Learning-based Detecting Approaches

Mining depression users based on TML mostly use features, i.e. numeric vectors that have been manually analyzed and extracted to represent the most important information of the predicted object (a user, a tweet, a posted picture, etc.).

Choudhury et al. [2] presented a pioneering work in this field of research. They provided a detailed feature engineering analysis process and a clear modeling approach by measuring the behavioral characteristics on Twitter users. Then, Wang et al. [16] implemented a sentiment analysis approach and proposed man-made rules by utilizing vocabulary to measure the depressive tendencies of tweets. Their work indicated that text-based features play a crucial role in online depression detection.

Deshpande et al. [12] proposed a representation learning method based on natural language processing (NLP) to model the text on Twitter. Different from the previously mentioned works [2], [16], their approach allows the classifier to automatically capture the latent features. After, Shen et al. [1] proposed an advanced detecting approach. They constructed a well-labeled depression detection dataset on Twitter, which had been widely used by subsequent researchers. Their proposed multimodal approach can effectively learn the latent and sparse representation of users' features.

Recently, more TML-based work has begun to emerge [14], [15]. Particularly, in the work of Ref. [15], they firstly introduced a neural network model for detecting depressed users on the OSN.

B. Deep Learning-based Detecting Approaches

DL-based detecting approaches can be used to automatically mine features. In particular, single-modal DL approaches are mainly oriented to the textual information. Researchers use NLP methods to embed text into high-dimensional continuous vectors to automatically mine text features, then construct a DL classification model for predicting depressed users. DL approaches based on multimodel ensemble learning and multimodal learning can comprehensively consider texts, pictures, videos, and other information from heterogeneous information sources. Specifically, multimodal-based approaches can be used to extract and fuse features from different sources to further improve model robustness [17]. Compared to TML, DL-based approaches are more flexible and efficient in processing various information.

Several DNN classifiers that have achieved significant performance in the NLP classification task were selected and evaluated by Orabi et al. [18]. Moreover, a pretrained Word2Vec [19] model was used to embed the tweet text. Their experimental results showed that compared to other recurrent network structures, CNN-based models performed better in the task of depression detection.

Then, Sadeque et al. [20] proposed a latency-weighted F1 metric and applied it in a novel sequential classifier based on

[2] https://github.com/aidenwang9867/Weibo-User-Depession-Detection-Dataset

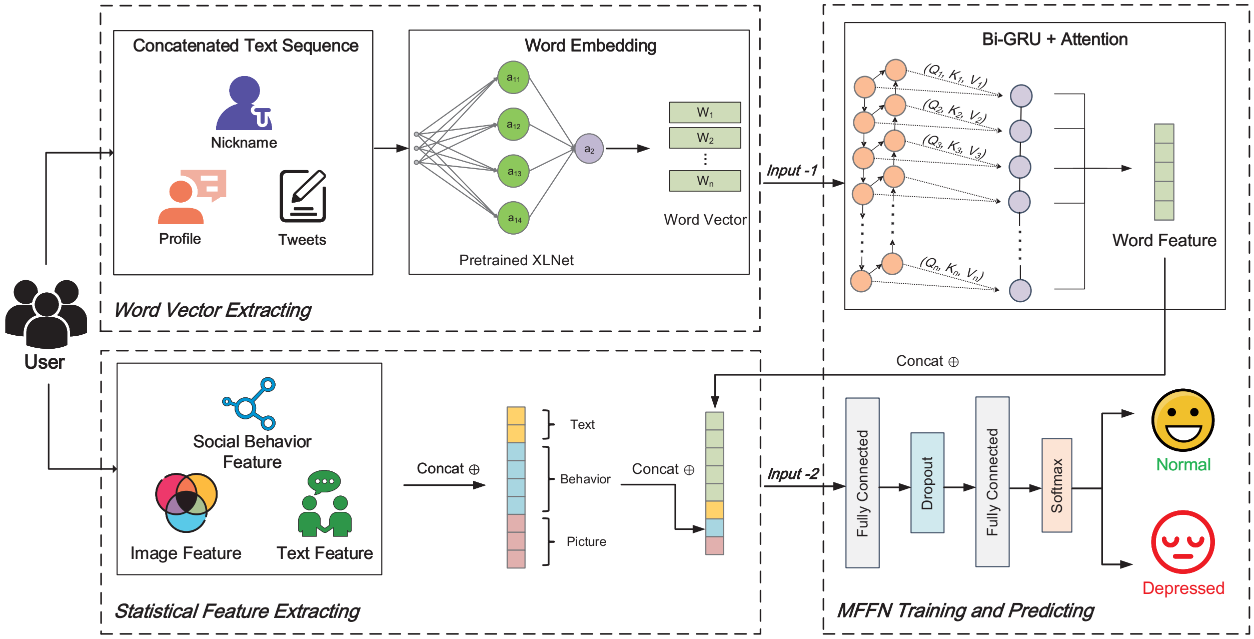


Fig. 1. The Framework of the Proposed Method

the Gated Recurrent Units (GRU). They treated all the text of tweets as a whole document and input it asynchronously to obtain classification results, which is named "post-by-post" strategy. This approach can scan and detect depressive tendencies of tweets more efficiently.

Based on their prior work [1], Shen et al. [3] discovered that the existing research on a specific OSN may be unsuitable and not universal for depression detection on other platforms. Thus, they proposed a cross-domain multimodal learning approach that can consider features of several aspects comprehensively and transfer the relevant information across heterogeneous domains.

Recently, more studies based on DL have been widely proposed. Gui et al. [4] further discussed the change of classification accuracy of the model under the different proportions of depressed users and pointed out that the highest accuracy can be achieved when the proportion of normal and depressed user samples is close to balance. Moreover, they implemented a reinforcement learning (RL) approach to further improve the performance of the model. Lin et al. [21] used a popular pretrained representation model, i.e. BERT [22], to embed word vectors. The neural network hidden layer output was extracted to fuse both text and picture-based features.

III. Methodology

To perform a more effective detection of depressed users on the OSN, we propose a novel framework, as shown in Fig. 1. This framework mainly consists of three parts.

i. *Word Vector Extracting.* This module is in charge of extracting the user's text information including the user's nickname, the profile, and the text of tweets, then concatenating them into a long text sequence. The sequence is input to the XLNet [23] pretrained model to obtain embedded word vectors.

ii. *Statistical Feature Extracting.* This module is responsible for extracting statistical features of users' tweet text, social behavior, and posted pictures.

iii. *Model Training and predicting.* This module implements a multimodal learning-based DNN classification model, namely Multimodal Feature Fusion Network (MFFN). MFFN has a network structure of Bi-GRU and the attention mechanism to reduce the dimensionality of word vectors received from the *Word Vector Extracting* module, i.e. *Input-1*. Then, the word feature is concatenated to the statistical features extracted from the *Statistical Feature Extracting* module, that is, *Input-2*. Finally, this module trains the network and gives classification results for normal and depressed users.

The following parts of this section will elaborate on the theoretical construction and the corresponding implementation methods, respectively.

A. User Data Collection and Labeling

1) Data collection: A user ID can be used to uniquely identify a user. With a user ID, the crawler can access the user's homepage and collect information from it. The user ID of depressed candidates was firstly crawled through the "Weibo Search" function provided by the Weibo official. Depression keywords such as "抑郁症" ("Depression" in English), "自杀" ("Suicide" in English), "痛苦" ("Pain" in English), and the late-night time period (0:00-6:00) were used as two search conditions. Another part of depressed candidates' user ID was crawled from the Weibo topic "抑郁症" ("Depression" in English). Then, more detailed user information fields were collected using the user ID.

Similarly, normal candidates were collected from four Weibo topics including “日常” (“Daily” in English), “正能量” (“Positive Energy” in English), “榜姐每日话题” (“Daily Topic” in English), “互动” (“Interaction” in English) to collect the user ID of normal candidates. Then, user information fields were collected to form the same data fields and structures as the depressed candidates. The information fields collected are shown in Fig. 2.

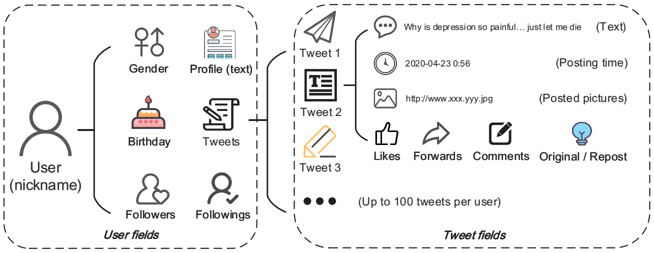


Fig. 2. The Data Structure of Candidates and WU3D (per user)

2) Data filtering and labeling: Automated scripts were implemented to filter out non-personal accounts by identifying the user’s *“account type”* field, including marketing accounts, official accounts, social bots. For normal candidates, we labeled the script-filtered candidates as normal users directly without further manual labeling process. For depressed users, we invited specialists of data labeling to complete the labeling. To ensure that the results are highly reliable, the labeled data has been reviewed twice by psychologists and psychiatrists.

Using the above steps, we constructed the target dataset, i.e. WU3D. Statistics of the candidates and WU3D are given in Table I. All the information of candidates was collected between March 2020 and May 2020, with a total number of over 200,000, including 125,479 depressed candidates and 65,913 normal candidates.

TABLE I
DATASET STATISTICS

Dataset	Category	User	Tweet	Picture
Candidates	Depressed	125,479	5,478,806	2,354,701
	Normal	65,913	4,927,904	3,631,537
	Total	191,392	10,406,710	5,986,238
WU3D	Depressed	10,325	408,797	160,481
	Normal	22,245	1,783,113	1,087,556
	Total	32,570	2,191,910	1,248,037

B. Statistical Feature Extracting

Previous studies have defined features that are effective for detecting depressed users, such as the proportion of late-night posted tweets, the number of tweets with negative sentiment, and the mean value of hue and saturation. In this part, we perform manual feature engineering in three aspects: the tweet text, social behavior, and posted pictures. Then, we summarize nine user-level features to perform depression detection, including four firstly proposed and two modified.[3] Descriptions of these features are shown in Table II.

Particularly, for **the proportion of negative emotional tweets**, We use the Text Sentiment Analysis API of the Baidu Smart Cloud Platform (version provided to medium and large enterprises with excellent Chinese text classification performance)[4] to label all the original tweets. The API returns three emotional labels: 0 for negative, 1 for neutral, and 2 for positive. We retain the negative emotions of label 0 and summarize labels 1 and 2 as a category of non-negative emotions. For **the posting frequency**, we take the earliest and latest release time as the interval, count the total number of posted tweets during the time, and then divide it by the total number of weeks to get the weekly frequency value. For **the standard deviation of posting time**, we convert the posting time of each tweet to a 24-hour format and calculate the standard deviation using all the tweets. For **the frequency of picture posting**, We divide the number of all the posted pictures by the number of all the original tweets.

C. Word Feature Extracting

1) Text sequence construction: For the construction of the long text sequence, we splice the user’s nickname, the profile (a self-introduction of the user), and the tweet text one by one. In the tweet text, all the original tweets and the reason filled by the user for reposts are used. Particularly, if the user does not fill in the reason, the text “转发微博” (“Repost” in English) will be automatically added as default. This default repost reason is not retained in the text sequence, since it does not express any opinions and feelings.

2) Word embedding: To effectively embed the text sequence constructed above and apply the word vector to the classification algorithm, several characteristics of the long text sequence are further discussed.

First, the sequence is strongly contextually linked. This link exists not only within a single tweet but also between the contexts of multiple tweets. For example, a user posts multiple tweets at different times with the content of depression diagnoses, depression onset, medication treatment, and inner distress. The integration of these information points is usually the key to judge whether a user is depressed.

Secondly, not all the tweets describe depression-related contents even for true depression users under real circumstances. That is to say, the ability to capturing text semantics such as “the user says that it has been diagnosed with depression” and “the user expresses a strong inclination of suicide” are critical for targeting depressed individuals.

Since XLNet combines the features of language models such as auto-regression and auto-encoding, it has resolved the problem that BERT [22] ignores the relationship between masked locations and can process longer sequences. Therefore, XLNet-Chinese-base[5] is used as the upstream word embedding

[3] We have conducted pre-experiments on our dataset. It proves that our modified features are more effective than the original ones.

[4] http://ai.baidu.com/tech/nlp/sentiment_classify

[5] https://github.com/ymcui/Chinese-XLNet

TABLE II
MANUALLY EXTRACTED USER FEATURES

Feature Group	Description	Source
Text	The proportion of negative emotional tweets.	**Firstly proposed in our work**
	The frequency of depression-related words.	Refs. [1]–[3], [16], [24], [25]
Social behavior	The proportion of original tweets.	Refs. [3], [16], [24]
	The proportion of late-night posting.	Refs. [1], [3], [16], [24]
	The posting frequency (per week).	**Firstly proposed in our work**
	The standard deviation of posting time.	**Firstly proposed in our work**
Picture	The frequency of picture posting.	**Firstly proposed in our work**
	The standard deviation of hue.	Refs. [1], [3], [24], modified in our work
	The standard deviation of saturation.	Refs. [1], [3], [24], modified in our work

model, then the MFFN is implemented as the downstream classification model to give predictive results of normal and depressed users.

IV. EXPERIMENT AND EVALUATION

A. Experiment Setup

1) Data cleaning: Considering that our construction and analysis of user word features are fully text-oriented, we have removed all the non-text contents in the user's nickname, the profile, and tweet text, to minimize the experimental bias and improve the efficiency of the model training.

2) Dataset slicing: In this part, WU3D is divided into four subsets: D_1, D_2, D_3, and D_4. All of the subsets are sampled using a fixed random seed without a crossover. D_1 is used for DNN model training and the ten-fold cross-validation of TML classifiers. Furthermore, D_2 is used as a fixed dataset for validation in each round of the neural network training. Finally, we evaluate the performance on D_3 and calculate the evaluation metrics. As a supplementary dataset, D_4 contains only 325 depressed users and 12,245 normal users, which will be only used in the last experiment of unbalanced training samples. Statistics of the sliced datasets are given in Table III:

TABLE III
DATASET SLICING STATISTICS

Dataset	Depressed			Normal		
	User	Tweet	Picture	**User**	Tweet	Picture
WU3D	10325	408797	160481	22245	1783113	1087556
D_1	8000	319115	125096	8,000	630064	355353
D_2	1000	37315	14991	1000	79417	46060
D_3	1000	38941	15211	1000	80066	45066
D_4	325	13426	5183	12245	993566	641077

3) Evaluation metrics: **True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN)** are commonly used to describe the number of classes predicted by models in classification tasks. Among them, TP represents the number of depressed users correctly predicted, TN represents the number of normal users correctly predicted, FP represents the number of normal users incorrectly predicted, and FN represents the number of depressed users incorrectly predicted. Based on the above four definitions, we can further define the metrics as:

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

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

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

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

4) Baseline statistical feature classifier: We select several popular classifiers from existing works as baselines to evaluate the classification contribution of the nine proposed statistical features.

- **LR:** Logistic Regression is a commonly used linear model and its classifier has good classification performance. It is often used as a baseline classifier in previous works [11], [26].
- **SVM:** The Support Vector Machine classifier applies the kernel learning method and is the most used baseline classifiers in previous studies [2], [12], [14], [15], [18], [20], [24]–[26].
- **RF:** Random Forest is an algorithm that integrates multiple classifiers through ensemble learning, which is also used widely in related works [15], [24], [27]. The basic unit of RF is the decision tree.
- **AB:** Adaptive Boosting is an ensemble learning algorithm that combines multiple simple classifiers. It has been used in Ref. [11].
- **GBDT:** Gradient Boosting Decision Tree uses an integrated additive model to continuously reduce the training residuals. GBDT is one of the algorithms in TML with excellent generalization abilities.
- **BP:** The fully connected network structure of the proposed MFFN, i.e. FC+Dropout+FC+Softmax.

5) Baseline word vector classification network: To verify the effectiveness of the word vector dimensionality reduction network in our proposed MFFN, i.e. the Bi-GRU with the attention layer, we implement several classifiers based on mainstream neural network structures as comparisons, to evaluate their classification performance on word vectors.

- **TCN:** The temporal convolutional network is a new algorithm for processing time series that reduces the serial processing complexity of RNNs [28].
- **Attention:** Attention is a mechanism proposed by Vaswani et al. [29] that can quickly filter out high-value

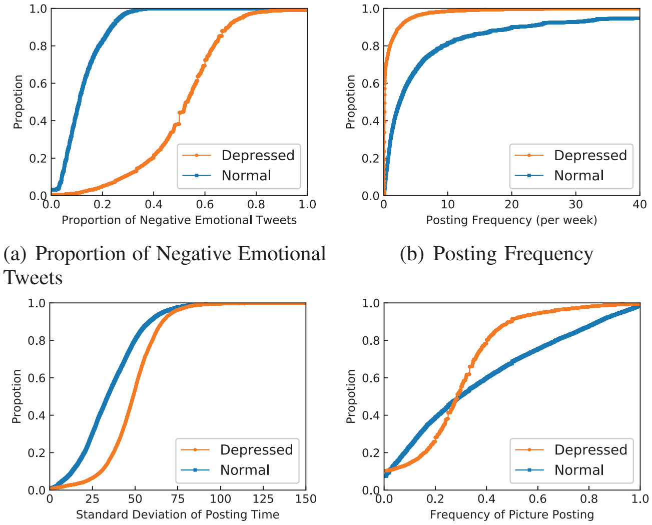
information from large amounts of information. Attention is popular in many fields such as machine translation and speech recognition.

- **CNN-1D:** One-dimensional convolutional neural networks are more widely used in natural language processing and have achieved good performance in the task of depression detection [11], [13], [15], [18], [24].
- **Bi-LSTM:** Long Short-Term Memory (LSTM) is a special kind of recurrent network. The bi-directional LSTM network splices two-way LSTMs together, which are more capable of handling time series data [18], [20], [21].
- **Bi-GRU:** GRU is a very popular variant of LSTM. The bi-directional GRU splices the two-way GRUs together [4], [20]. [28].
- **Bi-GRU with attention:** This structure is extracted from our proposed classification model MFFN, appended with a fully connected layer, and a softmax layer to obtain the classification results.

For the baseline statistical feature classifiers and word vector classification networks, we have run a series of pre-experiments on every classifier and selected the structures and parameters with the best classification performance. Each classifier will be represented by symbols of its main structures (e.g., Bi-GRU-based classifiers are referred to as Bi-GRU).

B. The Baseline Contribution of Different Features

1) Statistical Features: The baseline contribution metrics of statistical features are given in Table IV and shown in Fig. 4(a). The CDF curves of all the four features we proposed are plotted in Fig. 3. The results show the normal and depressed users' curves of all the features are highly differentiated. Both of the classification baseline and the CDF curves demonstrate that all the proposed features have significantly different distribution on depressed and normal users, thus can be used as valid classification features.



(a) Proportion of Negative Emotional Tweets

(b) Posting Frequency

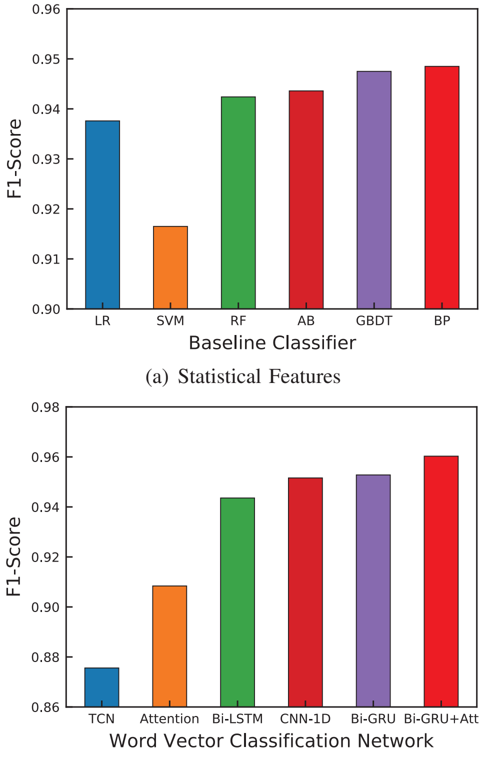
(c) Standard Deviation of Posting Time

(d) Frequency of Picture Posting

Fig. 3. CDF Curves of our Proposed Features

TABLE IV
STATISTICAL FEATURE CLASSIFIER BASELINES

Classifier	Accuracy	F1-Score	Precision	Recall
LR	0.9400	0.9376	0.9763	0.9019
SVM	0.9211	0.9165	0.9736	0.8657
RF	0.9445	0.9424	0.9804	0.9072
AB	0.9449	0.9436	0.9667	0.9215
GBDT	0.9491	0.9475	**0.9768**	0.9200
BP	**0.9492**	**0.9485**	0.9757	**0.9228**



(a) Statistical Features

(b) Word Features

Fig. 4. Baseline Performance of different Features

2) Word Features: The metrics of different word vector classification networks are shown in Table V and Fig. 4(b). Among them, the structure of Bi-GRU with Attention achieves the highest F1-Score. Although the performance of a single attention structure is relatively poor, by adding it to other structures, the classification performance can be further improved.

TABLE V
WORD VECTOR CLASSIFICATION NETWORK BASELINES

Classifier	Accuracy	F1	Preicsion	Recall
TCN	0.8752	0.8756	0.8951	0.8570
Attention	0.9088	0.9084	0.9388	0.8799
Bi-LSTM	0.9438	0.9436	0.9788	0.9108
CNN-1D	0.9517	0.9516	0.9772	0.9273
Bi-GRU	0.9530	0.9528	0.9793	0.9277
Bi-GRU+Attention	**0.9604**	**0.9603**	**0.9873**	**0.9347**

C. Performance of the Target Classifiers

To make it easier to compare other classifiers algorithm with the MFFN, in this section, the one-dimension output vector of

TABLE VI
Classification Performance of the Target Classifiers

Classifier	Accuracy	F1-Score	Precision	Recall
LR (Ensemble)	0.9562	0.9550	0.9824	0.9291
SVM (Ensemble)	0.9540	0.9526	0.9824	0.9245
RF (Ensemble)	0.9577	0.9566	0.9840	0.9306
AB (Ensemble)	0.9589	0.9579	0.9802	0.9366
GBDT (Ensemble)	0.9600	0.9588	0.9872	0.9321
MFFN (Proposed)	**0.9683**	**0.9685**	**0.9908**	**0.9472**

the Bi-GRU with attention network is extracted as the word feature. Then, it is concatenated to the statistical features as an integrated input to the baseline statistical feature classifiers.

Therefore, multimodel ensemble classifiers are constructed by both the word vector classification networks and the statistical feature classifiers. Table VI gives detailed metrics of all the target classifiers, while Fig. 5 visualizes them.

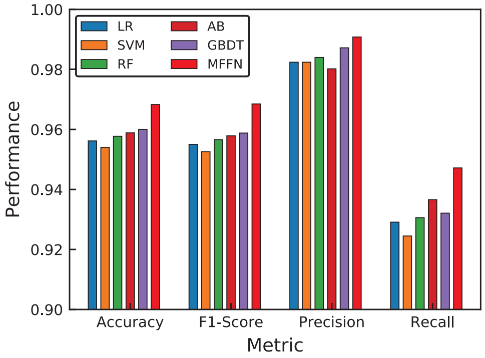


Fig. 5. Performance of Target Classifiers

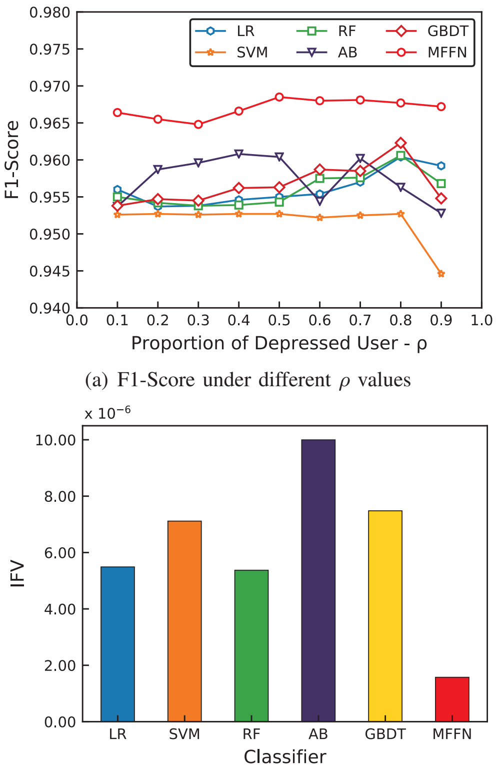
In the performance experiment of the target classifiers, each classifier reaches an F1-Score above 0.95. RF, AB, and GBDT have a certain degree of performance improvement due to the advantages of the boosting learning mode or jointly decision-making. Particularly, our proposed MFFN achieves the highest F1-Score value of 0.9685. It also obtains the highest value under all the other metrics. Compared to the multimodel ensemble approach, our implemented multimodal feature fusion approach has better classification performance.

D. Robustness of Unbalanced Training Samples

By changing the proportion of depressed user samples (denoted as ρ), we analyze the fluctuations of the F1-Score on the target classifiers to evaluate its robustness of training an unbalanced number of samples. Each classifier is treated as a group. For each group, we evaluate nine values of ρ from 0.1 to 0.9, with an interval of 0.1.

Here, we implement a new metric, namely the Intra-group F1-Score Variance (IFV), to calculate the variance of F1-Scores in each group. First, for each group, the mean value of the F1-Scores is calculated and represented by $\overline{X}_{IF}$. The number of ρ values taken in each group is denoted as λ. Then, the IFV metric is defined as:

$$IFV = \frac{1}{\lambda} \times \sum_{i=1}^{\lambda} (F1_i - \overline{X}_{IF})^2 \quad (5)$$



(a) F1-Score under different ρ values

(b) IFV of different classifiers

Fig. 6. Unbalanced Training Samples

The experimental result in Fig. 6(a) shows that due to the large base number of training samples, each classifier obtains an F1-Score higher than 0.94 at different ρ values. In the meantime, the IFV metrics shown in Fig. 6(b) further demonstrates that when the proportion of the two kinds of user samples changes, the F1-Score of MFFN fluctuates the least.

Furthermore, since the multimodal learning approach enables features from different sources to share the same network structure, it significantly reduces the loss of information caused by transfer learning that transfers the word feature into the TML classifiers as a part of the input. It also helps the classification model find the optimal solution faster and reach convergence. Therefore, it demonstrates that the multimodal approach has higher performance and robustness than the multimodel ensemble ones.

V. Conclusion

In this work, we propose a depression detection method based on multimodal feature fusion. Different forms of information from heterogeneous information sources are fused through feature engineering and text-based word embedding.

The experiments show that our proposed MFFN classification model has a better performance compared to several classifiers that are commonly used in existing studies. Moreover, the MFFN has better robustness when the proportion of training samples is unbalanced. Compared to multimodel detecting approaches, the multimodal approach allows information from heterogeneous sources to share the same network weight and to complement each other. In the meantime, this approach reduces the loss of information due to feature transfer in different models, thus enhances the classification capability and robustness of the models.

For future work, two directions can be explored. **(i)** The size of the dataset can be further expanded. We will build larger-scale datasets to achieve better generalization performance. **(ii)** User-level features based on the manual feature engineering can be further analyzed. We will continue to explore the characteristics and behavior patterns of depressed users to further propose effective feature solutions for early-stage user-level depression detection on the OSN.

Acknowledgment

The authors would like to thank anonymous reviewers for their helpful comments to improve the paper. The authors are also grateful to the anonymous data labelers and the psychologists and psychiatrists who provided advice.

References

[1] G. Shen, J. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T.-S. Chua, and W. Zhu, "Depression detection via harvesting social media: A multimodal dictionary learning solution," in *Proceedings of the 26th ACM International Joint Conference on Artificial Intelligence*, Melbourne, Australia, Aug 2017, pp. 3838–3844.

[2] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, Cambridge, MA, USA, Jul 2013, pp. 128–137.

[3] T. Shen, J. Jia, G. Shen, F. Feng, X. He, H. Luan, J. Tang, T. Tiropanis, T. S. Chua, and W. Hall, "Cross-domain depression detection via harvesting social media," in *Proceedings of the 27th ACM International Joint Conference on Artificial Intelligence*, Stockholm, Sweden, Jul 2018, pp. 1611–1617.

[4] T. Gui, L. Zhu, Q. Zhang, M. Peng, X. Zhou, K. Ding, and Z. Chen, "Cooperative multimodal approach to depression detection in twitter," in *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, Honolulu, HI, USA, Jan 2019, pp. 110–117.

[5] S. Li, Y. Wang, J. Xue, N. Zhao, and T. Zhu, "The impact of covid-19 epidemic declaration on psychological consequences: A study on active weibo users," *International Journal of Environmental Research and Public Health*, vol. 17, no. 6, p. 2032, Mar 2020.

[6] Y. Suhara, Y. Xu, and A. Pentland, "Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks," in *Proceedings of the 26th ACM International Conference on World Wide Web*, Perth, Australia, Apr 2017, pp. 715–724.

[7] H. Dibeklioğlu, Z. Hammal, and J. F. Cohn, "Dynamic multimodal measurement of depression severity using deep autoencoding," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 2, pp. 525–536, Mar 2017.

[8] Y. Zhu, Y. Shang, Z. Shao, and G. Guo, "Automated depression diagnosis based on deep networks to encode facial appearance and dynamics," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 578–584, Oct 2018.

[9] C.-T. Wu, D. G. Dillon, H.-C. Hsu, S. Huang, E. Barrick, and Y.-H. Liu, "Depression detection using relative eeg power induced by emotionally positive images and a conformal kernel support vector machine," *Applied Sciences*, vol. 8, no. 8, p. 1244, Jul 2018.

[10] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, Sep 2017.

[11] A. Cohan, B. Desmet, A. Yates, L. Soldaini, S. MacAvaney, and N. Goharian, "Smhd: a large-scale resource for exploring online language usage for multiple mental health conditions," in *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New-Mexico, USA, 2018, pp. 1485–1497.

[12] M. Deshpande and V. Rao, "Depression detection using emotion artificial intelligence," in *Proceedings of the 19th IEEE International Conference on Intelligent Sustainable Systems*, Palladam, Tirupur, India, Dec 2017, pp. 858–862.

[13] M. Trotzek, S. Koitka, and C. M. Friedrich, "Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 3, pp. 588–601, Mar 2020.

[14] H. S. AlSagri and M. Ykhlef, "Machine learning-based approach for depression detection in twitter using content and activity features," *arXiv preprint arXiv:2003.04763*, 2020.

[15] R. U. Mustafa, N. Ashraf, F. S. Ahmed, J. Ferzund, B. Shahzad, and A. Gelbukh, "A multiclass depression detection in social media based on sentiment analysis," in *Proceedings of the 17th IEEE International Conference on Information Technology—New Generations*. Las Vegas, NV, USA: Springer, Apr 2020, pp. 659–662.

[16] X. Wang, C. Zhang, Y. Ji, L. Sun, L. Wu, and Z. Bao, "A depression detection model based on sentiment analysis in micro-blog social network," in *Proceedings of the 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Gold Coast, QLD, Australia: Springer, Apr 2013, pp. 201–213.

[17] F. Huang, X. Zhang, J. Xu, Z. Zhao, and Z. Li, "Multimodal learning of social image representation by exploiting social relations," *IEEE Transactions on Cybernetics*, pp. 1–13, Mar 2019, early access.

[18] A. H. Orabi, P. Buddhitha, M. H. Orabi, and D. Inkpen, "Deep learning for depression detection of twitter users," in *Proceedings of the 5th Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, New Orleans, LA, USA, Jun 2018, pp. 88–97.

[19] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 26th ACM International Conference on Neural Information Processing Systems*, Lake Tahoe, NV, USA, Dec 2013, pp. 3111–3119.

[20] F. Sadeque, D. Xu, and S. Bethard, "Measuring the latency of depression detection in social media," in *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, Marina Del Rey, CA, USA, Feb 2018, pp. 495–503.

[21] C. Lin, P. Hu, H. Su, S. Li, J. Mei, J. Zhou, and H. Leung, "Sensemood: Depression detection on social media," in *Proceedings of the 28th ACM International Conference on Multimedia Retrieval*, Dublin, Ireland, Jun 2020, pp. 407–411.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[23] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in *Proceedings of the 33rd ACM International Conference on Neural Information Processing Systems*, Vancouver, Canada, Dec 2019, pp. 5753–5763.

[24] H. Lin, J. Jia, Q. Guo, Y. Xue, Q. Li, J. Huang, L. Cai, and L. Feng, "User-level psychological stress detection from social media using deep neural network," in *Proceedings of the 22nd ACM International Conference on Multimedia*, Orlando, FL, USA, Nov 2014, pp. 507–516.

[25] Q. Cheng, T. M. Li, C.-L. Kwok, T. Zhu, and P. S. Yip, "Assessing suicide risk and emotional distress in chinese social media: A text mining and machine learning study," *Journal of Medical Internet Research*, vol. 19, no. 7, p. e243, Jul 2017.

[26] I. Sekulić and M. Strube, "Adapting deep learning methods for mental health prediction on social media," *arXiv preprint arXiv:2003.07634*, 2020.

[27] S. Balani and M. De Choudhury, "Detecting and characterizing mental health related self-disclosure in social media," in *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, Seoul, Republic of Korea, Apr 2015, pp. 1373–1378.

[28] A. Pandey and D. Wang, "Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *Proceedings of the 44th IEEE International Conference on Acoustics, Speech and Signal Processing*, Brighton, UK, May 2019, pp. 6875–6879.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st ACM International Conference on Neural Information Processing Systems*, Long Beach, CA, USA, Dec 2017, pp. 5998–6008.