Depression Detection Method Using Multimodal Social Media Data: An Integrative Review

Jiawen Ma

School of Information Science and Engineering, Yanshan University

majiawen@stumail.ysu.edu.cn

Abstract. An increasing number of people are suffering from depression due to rising chronic stress levels. With the advent of Web 2.0, individuals are more inclined to express their emotions on social media, offering new opportunities for depression prediction. Researchers have developed various single-modal methods for early-stage depression prediction. Recently, multimodal social media data has been utilized to enhance the accuracy of depression detection methods. These methods primarily extract multidimensional information such as text, language, and images from social media users, integrating these diverse modes to assess the risk or severity of depression. This approach significantly improves the precision of depression prediction. However, the research is still in its early stages, with challenges such as limited datasets and many areas requiring further improvement. To aid researchers in better understanding and refining multimodal approaches, we conducted a review that summarizes emerging research directions in using multimodal techniques for depression prediction on social media. Additionally, this review compares different depression detection methods, datasets, and the various modalities used in multimodal approaches, analyzing their strengths and limitations. Finally, it offers suggestions for future research.

Keywords: depression detection, multimodal social media data, deep learning, feature fusion.

1. Introduction

The number of people suffering from depression is increasing worldwide, seriously affecting the health and well-being of individuals. According to the World Health Organization (WHO), 5 percent of adults globally suffer from depression [1]. However, due to limitations in medical resources, the diversity of symptoms, social stigma, and other factors, many patients with depression are not diagnosed at an early stage. Traditional approaches rely on clinical consultations or questionnaires, such as the Patient Health Questionnaire (PHQ-9) and the Montgomery-Åsberg Depression Rating Scale (MADRS), which are commonly used to assess depression [2]. In recent years, social media use has been on the rise, and the emotions expressed by individuals with depression on these platforms often appear more authentic than those captured by traditional methods, like questionnaires. Therefore, the abundance of real-time, diverse data available on social media offers new opportunities for predicting depression. By collecting data from a single modality on social media (e.g., text, images, audio, or video), previous researchers have analyzed these datasets to assess user characteristics and predict their likelihood of depression. For instance, M. Trotzek et al. employed machine learning models based on social media posts to address the problem of early depression detection [3]. However, single-modal methods, which rely on one type of data, often result in insufficient predictive accuracy and struggle to account for individual differences. To achieve more accurate and effective predictions of depression, researchers have begun exploring multimodal approaches to address these limitations. Multimodal models have indeed been shown to predict depression more effectively than single-modal methods. By effectively integrating data from multiple modes, such as text, images, and audio, researchers can extract more accurate features, leading to improved classification and analysis. As a result, the accuracy of depression predictions is significantly enhanced. However, multimodal approaches still face several challenges, including a lack of publicly available datasets, difficulty in selecting features from different data modalities, and issues with model overfitting. To help researchers better understand and refine multimodal methods, we analyze recent research on both single-modal and multimodal approaches for depression prediction and provide recommendations for future research.

2. Literature Review

2.1. Depression Detection Methods Using Single-Modal Social Media Data

Earlier studies aimed to predict depression using single-modal techniques. Researchers collected text, image, audio, or video information from social media users and analyzed their depression levels by extracting features from a single modality. Single-modal research primarily employed machine learning methods, which included techniques such as linear regression, decision trees, K-nearest neighbors, Naive Bayes, random forests, and support vector machines. Many studies focused on text analysis. For instance, R. Chiong et al. demonstrated the effectiveness of using machine learning methods in their research [4]. However, their approach relied on supervised machine learning classifiers, which require labeled datasets for training. This limitation could be addressed in the future by introducing unsupervised classifiers. Later, Y. Wu et al. introduced a novel approach by defining the basic components of personalized information within the domain of text-based depression prediction and proposing the Personalized Information Embedded (PIE) model [5]. This model emphasized the extraction of personalized information, thereby improving the accuracy of predictions.

As technology advanced, researchers began to apply deep learning methods. Compared to traditional machine learning, the primary advantage of deep learning is its ability to achieve higher accuracy when processing large-scale data. For example, A. Pérez et al. investigated the use of neural language models to detect depression by designing a classification framework [6]. Additionally, J. Singh et al. proposed an embedded long short-term memory (LSTM)-based scheme that utilized natural language processing technology [7]. This embedded neural network scheme effectively predicted users' emotional states. Furthermore, V. Tejaswini et al. introduced a novel hybrid deep learning neural network model combining FastText, convolutional neural networks, and LSTM [8] [9]. This model integrated NLP with deep learning to represent both global and local features of depression effectively.

Although single-modal methods have produced promising results, they also have limitations, such as unstable classification performance and incomplete feature extraction, leading to inaccurate depression predictions. As a result, researchers have turned to multimodal approaches to improve the prediction of depression.

2.2. Depression Detection Methods Using Multimodal Social Media Data

The multimodal research method analyzes the characteristics of different modalities more comprehensively, resulting in more accurate predictions. Multimodal prediction primarily relies on multimodal fusion methods, which can be broadly categorized into three types: feature-level fusion (early fusion), decision-level fusion (late fusion), and model-level fusion. Most current research focuses on feature-level fusion.

The fusion of text and images has shown promising results. For example, T. Gui et al. proposed a new cooperative multi-agent model [10] that automatically selects relevant text and images from users' historical posts. These selected posts effectively indicate a user's depression, and the model demonstrated strong and stable performance in real-world scenarios. Additionally, C. Lin et al. designed a system called SenseMood [11], which used a deep visual-textual multimodal learning approach to reveal users' psychological states on social networks by analyzing the pictures and text they posted. This system significantly improved depression detection performance on social networks. Some studies also incorporate features from text, video, and audio. For instance, M. Li et al. proposed a Multimodal Fusion Module (MFM) and introduced the IISFD method to enhance comparative learning between samples [12]. There are also other feature fusion methods, such as the one proposed by H. Fan et al., which integrates video, audio, and rPPG signals using a Transformer-based multimodal feature enhancement network (TMFE). This approach was the first to combine rPPG signals with video and audio modes for multimodal depression detection, proving helpful in detecting depression [13]. Z. Li et al. proposed a Multimodal Hierarchical Attention (MHA) model for depression detection on social media [14]. Using the depression detection dataset from Weibo, this model demonstrated good classification performance and accurately identified depressed users on social media.

Feature fusion has also been combined with deep learning techniques. A. Malhotra and R. Jindal proposed a real-time, deep learning-based system for sentiment analysis of users' multimodal social media posts. The system extracts features from images, text, and videos on a user's social media through multiple patterns to detect episodes of depression, suicidal ideation, or self-harming behavior. They used deep learning techniques to detect depression and suicidal tendencies in real-time by analyzing multimodal user-generated content [15]. Additionally, P. Mann and A. Paes, along with their colleagues, detected depression using multimodal social media data. They compared deep learning models to feature-engineered models derived from Instagram images and their captions. By creating three types of deep learning and feature-engineering models—text-only, image-only, and text+image fusion—they achieved the best prediction results using ELMo and ResNet-34 classifiers to combine features. This approach could assist in large-scale depression screening [16]. Furthermore, H. Zogan et al. proposed a new computational framework for automated depression detection. The framework automatically summarizes a user's post history to select relevant user-generated content. This content is then processed through a novel deep learning framework, consisting of a CNN-GRU model, which outperformed existing robust baselines [17]. H. Zhang et al. introduced the neural network hybrid model MTDD to identify depressive tendencies [18]. Based on social platform data, this deep neural network hybrid model utilized multimodal features to learn the vector representation of depression-prone text, proving effective in detecting depressive tendencies.

Using feature fusion methods to construct multimodal sentiment dictionary models has also shown promising results. M. P. M. Tadlagi et al. proposed a multimodal depressive dictionary learning method to detect depressed users on Twitter [19]. This method was developed using benchmark depression and non-depression datasets, along with well-defined discriminative depression-oriented feature groups. It revealed potential differences in online behavior between depressed and non-depressed users on social media.

Recent studies have shown that hybrid fusion methods yield better results than simple early or late fusion approaches. For example, X. Zhang et al. proposed a novel trilateral bimodal encoding model (MEN), attentional decision fusion (ADF), and a feature extraction fusion strategy [20]. This method combined early intra-modal fusion with late inter-modal fusion, representing a hybrid fusion approach.

Although the accuracy of depression prediction has significantly improved with the use of multimodal methods, there are still some limitations: public datasets suitable for research remain insufficient, there is a risk of disclosing the privacy of depression patients, and the constructed models may suffer from overfitting. Therefore, the multimodal approach still has considerable room for improvement, offering opportunities for further exploration by researchers.

3. Discussion

3.1. General Process of Multimodal Depression Detection Methods

Based on recent literature, we summarize the general steps of multimodal prediction methods, as shown in Fig. 1:

(1) Multimodal data acquisition: Collecting text, image, video, audio, and other data from users on social media.

(2) Data cleaning and preprocessing: Cleaning the collected data to remove noise and outliers, and preprocessing it to facilitate subsequent analysis.

(3) Single-mode feature extraction: Using the appropriate techniques for each mode to extract relevant features.

(4) Multimodal feature fusion: Fusing the extracted features, which can involve simple feature fusion or deep learning-based methods.

(5) Model construction and training: Selecting an appropriate model and training it using the datasets.

(6) Model evaluation and optimization: Evaluating the model's performance using suitable metrics and further optimizing it.

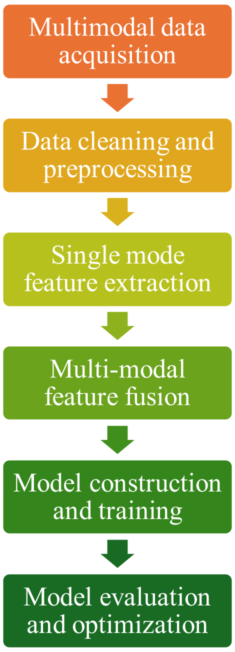
Multimodal data acquisition
Data cleaning and preprocessing
Single mode feature extraction
Multi-modal feature fusion
Model construction and training
Model evaluation and optimization

Figure 1

3.2. Comparative Study of Multimodal and Single-modal Methods

Table 1. Comparison of single-modal and multimodal datasets and techniques

	Authors	Year	Datasets	Techniques
Single-modal	R. Chiong et al. [4]	2021	Labeled Twitter datasets, depression-class-only datasets from Facebook, Reddit, and an electronic diary	BOW, LR, LSVM, MLP, DT model
	A. Pérez et al. [6]	2022	Reddit data	Word2Vec, Question Answering (QA) model, BERT
	J. Singh et al. [7]	2022	Reddit depression dataset	NLTK library, GloVe, LSTM model
	Y. Wu et al. [5]	2024	Intervention dataset, multi-stimulus depression dataset	BERT, VAE, contrastive learning techniques

Table 1. (continued).

			(MIDD), Chinese Multimodal Depression Corpus (CMDC)	
	X. Xu et al. [21]	2024	CNRAC and CS-NRAC datasets	Voice Activity Detection (VAD), Librosa, OpenSMILE libraries, Fast Fourier Transform, LSTM, DWA, MFCC
	V. Tejaswini et al. [9]	2024	Reddit posts and tweets from Twitter	fasttex, LSTM, CNN
Multi-modal	T. Gui et al. [10]	2019	Textual Depression Dataset, Multimodal Depression Dataset D2	GRU, VGG-Net, Naive Bayes, multiple social networking learning, Wasserstein dictionary learning, multimodal depressive dictionary learning
	P. Mann et al. [16]	2020	Instagram data	ResNet deep network, ResNeXt network, ImageNet1k, PyTorch framework, Bag of Words, FastText, ELMo, FC (Fully Connected layer)
	A. Malhotra et al. [15]	2020	Facebook, Instagram, Twitter	VGG-16 network, Word2Vec, Skip Gram model or CBOW model, Faster R-CNN, 3D convolution layers (conv3d), FC
	C. Lin et al. [11]	2020	Tweets from users with/without depression on Twitter	CNN, BERT
	M. P. M. Tadlagi et al. [19]	2022	Depression dataset, non-depression dataset, depression-candidate dataset from Twitter	Porter Stemmer, Word2Vec, LIWC, Latent Dirichlet Allocation model, multimodal depressive dictionary learning model
	Z. Li et al. [14]	2023	Sina Weibo data	ResNet-18, PyTorch 1.10 framework, NVIDIA A100 Tensor Core GPU, TextCNN, DPCNN, FastText, BERT, Transformer
	H. Zhang et al. [18]	2023	Social platform data	Word2Vec, TF-IDF model, emotional dictionary, CNN model, BiLSTM network
	A. Das et al. [22]	2024	DAIC-WOZ, MODMA, RAVDESS	MFCC, batch normalization, convolution 2D, Leaky ReLU, Fast Fourier Transform (FFT), PyTorch
	X. Zhang et al. [20]	2024	AVEC2013, AVEC2014	CNN, BiLSTM

Table 1 shows that datasets used in single-modal methods typically contain only one type of data, and the techniques applied are relatively simple. However, relying on a single source of information may not fully capture the patient's true condition. In contrast, multimodal datasets integrate multiple types of data and employ more advanced techniques, which significantly enhances prediction accuracy.

3.3. Comparative Study of the Modes and Techniques Used in Multimodal Methods

Table 2. Comparison of Modes and Techniques Used in Multimodal Approaches

Authors	Year	Modes	Techniques
T. Gui et al. [10]	2019	Text, images	GRU, VGG-Net, Naive Bayes, multiple social networking learning, Wasserstein dictionary learning, multimodal depressive dictionary learning
P. Mann et al. [16]	2020	Text, images	ResNet deep network, ResNeXt network, ImageNet1k, PyTorch framework, Bag of Words, FastText, ELMo, Fully Connected (FC) layer
A. Malhotra et al. [15]	2020	Text, images, videos	VGG-16 network, Word2Vec, Skip-gram model or CBOW model, Faster-RCNN, 3D convolution layers (conv3d), FC layer
C. Lin et al. [11]	2020	Text, images	CNN, BERT
M. P. M. Tadlagi et al. [19]	2022	Social network features, user profile features, visual features, emotional features, topic-level features, domain-specific features	Porter Stemmer, Word2Vec model, LIWC, Latent Dirichlet Allocation (LDA) model, multimodal depressive dictionary learning model
Z. Li et al. [14]	2023	Text, images, auxiliary information	ResNet-18, PyTorch 1.10 framework, single NVIDIA A100 Tensor Core GPU, TextCNN model, DPCNN, FastText, BERT, Transformer
H. Zhang et al. [18]	2023	Text features, semantic features, domain knowledge	Word2Vec, TF-IDF model, emotional dictionary, CNN model, BiLSTM network
A. Das et al. [22]	2024	Speech, audio, MFCC features	MFCC, batch normalization, 2D convolution, Leaky Rectified Linear Unit (Leaky ReLU), Fast Fourier Transform (FFT), PyTorch
M. Li et al. [12]	2024	Vision, audio, text	Short-time Fourier transform, BERT model, CMF, MDD, PyTorch
H. Fan et al. [13]	2024	Video, audio, rPPG	Transformer-based multimodal feature enhancement, graph fusion network, Multi-task Cascade Convolutional Neural Networks

According to the comparison shown in Table 2, it is evident that multimodal feature fusion is the most commonly used method among multimodal approaches. The most frequently used modalities are text, images, and videos. Feature extraction techniques predominantly include the TF-IDF model, Word2Vec model, GloVe, BERT model, SVM, CNN, and others. For feature fusion, commonly used techniques are FC (Fully Connected layer), LSTM, and CNN. Multimodal feature fusion methods can range from simple feature fusion to combinations of machine learning and deep learning models.

4. Conclusion

By analyzing and comparing single-modal and multimodal depression detection methods on social media, it is apparent that while single-modal depression prediction has certain application value, it is limited by issues related to information comprehensiveness and prediction accuracy. Multimodal depression prediction, on the other hand, provides more comprehensive and rich information by combining multiple data types, thereby improving prediction accuracy. However, the multimodal approach still faces challenges, including insufficient publicly available datasets suitable for research, complexities in data feature extraction and fusion, and overfitting in constructed models. To enhance the predictive performance of multimodal methods, researchers should focus on developing efficient multimodal data acquisition techniques, optimizing the fusion methods for multimodal data, and addressing individual differences between samples. The application of multimodal prediction for detecting depression in patients on social media holds significant potential. Researchers should foster interdisciplinary collaboration among psychology, computer science, neuroscience, and other fields, and increase clinical testing to contribute to the early detection of depression.

References

[1] W. H. O. (WHO), “Depressive disorder (depression).” Accessed: Sep. 05, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression

[2] H. K. Kim et al., “Relationship between Patient Health Questionnaire (PHQ-9) and Montgomery-Asberg Depression Rating Scale (MADRS) total scores in older adults with major depressive disorder: An analysis of the OPTIMUM clinical trial,” J. Affect. Disord., vol. 361, no. December 2023, pp. 651–658, 2024, doi: 10.1016/j.jad.2024.06.068.

[3] M. Trotzek, S. Koitka, and C. M. Friedrich, “Utilizing Neural Networks and Linguistic Metadata for Early Detection of Depression Indications in Text Sequences,” IEEE Trans. Knowl. Data Eng., vol. 32, no. 3, pp. 588–601, 2020, doi: 10.1109/TKDE.2018.2885515.

[4] R. Chiong, G. S. Budhi, S. Dhakal, and F. Chiong, “A textual-based featuring approach for depression detection using machine learning classifiers and social media texts,” Comput. Biol. Med., vol. 135, p. 104499, 2021, doi: 10.1016/j.compbiomed.2021.104499.

[5] Y. Wu et al., “PIE: A Personalized Information Embedded model for text-based depressi on detection,” Inf. Process. Manag., vol. 61, no. 6, p. 103830, 2024, doi: 10.1016/j.ip m.2024.103830.

[6] A. Pérez, J. Parapar, and Á. Barreiro, “Automatic depression score estimation with word embedding models,” Artif. Intell. Med., vol. 132, no. June, p. 102380, 2022, doi: 10.1016/j.artmed.2022.102380.

[7] J. Singh, M. Wazid, D. P. Singh, and S. Pundir, “An embedded LSTM based scheme for depression detection and analysis,” Procedia Comput. Sci., vol. 215, pp. 166–175, 2022, doi: 10.1016/j.procs.2022.12.019.

[8] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, “Bag of tricks for efficient text classification,” 15th Conf. Eur. Chapter Assoc. Comput. Linguist. EACL 2017 - Proc. Conf., vol. 2, pp. 427–431, 2017, doi: 10.18653/v1/e17-2068.

[9] V. Tejaswini, K. S. Babu, and B. Sahoo, “Depression Detection from Social Media Text Analysis using Natural Language Processing Techniques and Hybrid Deep Learning Model,” ACM Trans. Asian Low-Resource Lang. Inf. Process., vol. 23, no. 1, 2024, doi: 10.1145/3569580.

[10] T. Gui et al., “Cooperative multimodal approach to depression detection in twitter,” 33rd AAAI Conf. Artif. Intell. AAAI 2019, 31st Innov. Appl. Artif. Intell. Conf. IAAI 2 019 9th AAAI Symp. Educ. Adv. Artif. Intell. EAAI 2019, pp. 110–117, 2019, doi: 10.1609/aaai.v33i01.3301110.

[11] C. Lin et al., “SenseMood: Depression detection on social media,” ICMR 2020 - Proc. 2020 Int. Conf. Multimed. Retr., pp. 407–411, 2020, doi: 10.1145/3372278.3391932.

[12] M. Li, Y. Wei, Y. Zhu, S. Wei, and B. Wu, “Enhancing multimodal depression detection with intra- and inter-sample contrastive learning,” Inf. Sci. (Ny)., vol. 684, no. July, p. 121282, 2024, doi: 10.1016/j.ins.2024.121282.

[13] H. Fan et al., “Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals,” Inf. Fusion, vol. 104, no. July 2023, p. 102161, 2024, doi: 10.1016/j.inffus.2023.102161.

[14] Z. Li, Z. An, W. Cheng, J. Zhou, F. Zheng, and B. Hu, “MHA: a multimodal hierarchical attention model for depression detection in social media,” Heal. Inf. Sci. Syst., vol. 11, no. 1, pp. 1–13, 2023, doi: 10.1007/s13755-022-00197-5.

[15] A. Malhotra and R. Jindal, “Multimodal deep learning based framework for detecting depression and suicidal behaviour by affective analysis of social media posts,” EAI Endorsed Trans. Pervasive Heal. Technol., vol. 6, no. 21, pp. 1–9, 2020, doi: 10.4108/eai.13-7-2018.164259.

[16] P. Mann, A. Paes, and E. H. Matsushima, “See and read: Detecting depression symptoms in higher education students using multimodal social media data,” Proc. 14th Int. AAAI Conf. Web Soc. Media, ICWSM 2020, no. Icwsm, pp. 440–451, 2020, doi: 10.1609/icwsm.v14i1.7313.

[17] H. Zogan, I. Razzak, S. Jameel, and G. Xu, “DepressionNet: Learning Multi-modalities with User Post Summarization for Depression Detection on Social Media,” SIGIR 2021 - Proc. 44th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr., pp. 133–142, 2021, doi: 10.1145/3404835.3462938.

[18] H. Zhang, H. Wang, S. Han, W. Li, and L. Zhuang, “Detecting depression tendency with multimodal features,” Comput. Methods Programs Biomed., vol. 240, no. January 2022, p. 107702, 2023, doi: 10.1016/j.cmpb.2023.107702.

[19] M. P. M. Tadlagi, M. V. P. Deshpande, M. A. A. Gaffar Chanda, M. P. R. Kakade, and D. K. K. S. Liyakat, “Depression Detection,” J. Ment. Heal. Issues Behav., no. 26, pp. 1–7, 2022, doi: 10.55529/jmhib.26.1.7.

[20] X. Zhang, B. Li, and G. Qi, “A novel multimodal depression diagnosis approach utilizing a new hybrid fusion method,” Biomed. Signal Process. Control, vol. 96, no. PA, p. 106552, 2024, doi: 10.1016/j.bspc.2024.106552.

[21] X. Xu, Y. Wang, X. Wei, F. Wang, and X. Zhang, “Attention-based acoustic feature fusion network for depression detection,” Neurocomputing, vol. 601, no. July, p. 128209, 2024, doi: 10.1016/j.neucom.2024.128209.

[22] A. K. Das and R. Naskar, “A deep learning model for depression detection based on MFCC and CNN generated spectrogram features,” Biomed. Signal Process. Control, vol. 90, no. November 2023, p. 105898, 2024, doi: 10.1016/j.bspc.2023.105898.