Dean&Francis
ISSN 2959-6157

Monitoring of Mental Illnesses Based on Natural Language Processing

Benwen Xu

School of Cyber Science and Engineering, Huazhong University of Science and Technology, Wuhan, Hubei, China
*Corresponding author: u202211956@hust.edu.cn

Abstract:
With the development of the Internet, more people are beginning to share their feelings and thoughts online. At the same time, the mental health issues of modern people are becoming increasingly prominent, gradually becoming a major source of suffering in people's lives and affecting the health and well-being of society. Mental illness is a complex, multifactorial disease associated with individual risk factors as well as various socioeconomic and clinical correlations. Therefore, the scientific community has developed a series of mental illness detection methods, among which natural language processing (NLP) plays a crucial role in mental illness detection. However, despite its rapid development in mental illness detection, the current detection methods still have some limitations. It cannot judge the authenticity of online information. Meanwhile, most detection models are only based on keyword analysis, resulting in overly simplistic evaluation criteria and a tendency to lead to some prejudgments. This paper mainly discusses the current methods and limitations of NLP models in mental illness detection, as well as possible solutions, aiming to provide some clues and assistance for future research.

Keywords: NLP; Mental illness; Prediction; Depression.

1. Introduction

As a category of complex health issues, mental illnesses continue to highlight their depth and breadth of impact on a global scale. They directly lead to a decline in patients' social adaptability, impairment of quality of life, and often accompany a chain reaction of physical health, forming a pathological cycle of interaction between the body and the mind. The research, prevention, and treatment of mental illnesses are not only medical problems but have also become core issues in the fields of public health and social development, urgently requiring interdisciplinary integrated response strategies.

With the rapid development of the Internet, it has evolved from the early stage of one-way information dissemination to multi-directional interaction. The continuous lowering of technical thresholds and the increasing improvement of platform ecosystems have greatly unleashed individuals' willingness to express themselves. Nowadays, people can conveniently make public their views on social events, insights into life experiences, and even thoughts on niche

ISSN 2959-6157

fields through diverse forms such as text, voice, and images.

In modern society, people generally endure high-intensity and multi-source pressures. When exposed to a state of accumulated stress for a long time, an individual's psychological regulation system gradually becomes imbalanced, emotional stability declines, and the tolerance threshold for negative stimuli decreases. This continuous psychological exhaustion weakens the mental defense line, significantly increasing the risk of various mental illnesses. Under the influence of long-term stress, it is more likely to develop into persistent psychological disorders, causing substantial damage to physical and mental health[1].

The development of Natural Language Processing (NLP) has advanced progressively through iterations of technical paradigms. In its early stages, it centered on manual rules, parsing language through preset grammatical logic, which struggled to cope with the flexibility of language. It then shifted to statistical methods, relying on large-scale data to train probabilistic models, which made tasks like machine translation practical. However, this approach depended on manually designed features and could hardly reach deep semantic levels. The introduction of neural networks brought a breakthrough: word embedding technology converted vocabulary into computable vectors, recurrent neural networks handled sequence dependencies, and the Transformer architecture, with its self-attention mechanism, broke through the bottleneck in processing long texts. The pre-training model was further upgraded, adapting to specific tasks through fine-tuning, and models such as BERT and GPT continuously refreshed performance records. Nowadays, large models are driving NLP toward general artificial intelligence, demonstrating language understanding and generation capabilities close to those of humans in complex scenarios.

The following sections will be presented in sequence. The first subsection describes the previous detection methods for major mental illnesses and their achieved effects; the second subsection discusses some of the current challenges and solutions; and finally, a prospect and summary are provided.

2. Traditional Methods for Detecting Mental Illnesses

Nowadays, as mental illnesses become increasingly severe, various detection methods have emerged, evolving from early traditional approaches to today's intelligent ones. Traditional diagnostic methods for mental illnesses, such as scale assessments and in-person consultations with doctors, have limitations including strong subjectivity, high requirements for patient cooperation, and low efficiency. However, with the popularization of the Internet, massive amounts of text data related to mental health have accumulated through channels like social media and medical records. Against this backdrop, NLP technology, with its powerful text analysis capabilities, has been widely applied in the field of mental illness detection. A large number of NLP models continue to emerge, dedicated to accurately mining clues of mental illnesses from texts to assist in diagnosis and condition monitoring. The text will mainly introduce five of the most common and effective NLP models for detecting mental illnesses: BERT and its variant models (as baseline methods), n-gram-based linear networks, text convolutional neural networks (CNN), and hierarchical attention networks (HAN)[2].

2.1 Detection of Mental Illnesses Based on BERT

In the detection and rehabilitation assessment of severe mental illnesses such as schizophrenia and major depressive disorder, BERT has become a key technology due to its strong ability to process unstructured texts. Traditional assessments rely on manual interviews, which suffer from issues such as single-source data and subjectivity. In contrast, through bidirectional contextual modeling, BERT can parse doctors' notes and patients' follow-up texts, capture implicit semantics like "low mood" and "refusal to take medication," and extract in-depth disease-related features.

In practical applications, BERT is often combined with structured data processing models such as TabNet to construct detection systems through multimodal fusion. For example, in the model proposed by Yang et al. (2024), BERT processes text data and collaboratively completes the detection of four indicators including referral risk and medication adherence, with an accuracy rate exceeding 78% (medication adherence reaching 94.3%). Moreover, it achieves decision interpretability through the attention mechanism, facilitating clinical personalized treatment and decision optimization [3].

This detection method prioritizes BERT variants adapted to tweet scenarios, uniformly adopting a basic architecture consisting of 12 Transformer blocks and 110 million parameters. The core logic behind its selection primarily centers on "semantic adaptability to tweets". It focuses on testing vinai/bertweet-base, while also incorporating bert-base-cased/uncased and bert-base-multilingual-cased. Its architectural design involves connecting a fully connected layer in series at the output end of the BERT encoder, forming a "pre-training - fine-tuning" closed loop. In the pre-training phase, it relies on BERT's existing semantic capabilities; in the fine-tuning phase, labeled tweet data is used to optimize the parameters of the fully connected layer. The input text only needs to undergo character-level processing without manual feature extraction, thereby lowering the technical threshold.

In terms of data processing, data collection and annotation are carried out in two main phases. In the first phase, Twitter tweets from April to August 2022 were collected, and 3,424 valid data entries were finally screened. These entries cover 11 items across 5 dimensions of mental health literacy (M/R/S/HA/HE), with each item corresponding to 5 intensity scores ranging from -2 to 2 (to quantify the degree of semantic association). It was ensured that each intensity level of each item had at least 60 samples. Subsequently, personnel with mental health backgrounds (such as clinical psychology doctors and psychiatric nursing doctors) conducted independent double annotations, and the annotation consistency reached a Pearson correlation coefficient of 0.8011, providing a reliable benchmark for model evaluation.

The data was divided into training and test sets using a stratified splitting method (4:1), meaning samples of each intensity level for each item were allocated as "4 parts for training and 1 part for testing." This avoids data distribution bias caused by a high proportion of "irrelevant samples (intensity 0)" and ensures fair model evaluation. The core task of the experiment is to automatically classify tweets into 11 mental health literacy items and output intensity scores, which essentially involves detecting semantic information related to mental illnesses implied in the tweets. The Macro F1 score is used as the core indicator (target ≥ 0.8, matching the consistency of manual annotations, with a focus on measuring the classification balance of each intensity score), supplemented by reference to Micro F1 (overall accuracy). Meanwhile, the training time for individual items and the prediction time for single tweets are recorded to balance performance and practical application efficiency [4].

In terms of overall performance, among the four BERT variants, vinai/bertweet-base achieved the best results. However, the Macro F1 scores of all variants failed to reach the target of 0.8, ranging from 0.33 to 0.78 overall. Regarding different items, those with explicit semantics (such as "Maintaining positive mental health (M5)" and "Recognition of mental illnesses (R2)") performed better (with Macro F1 exceeding 0.7), while "Help-seeking attitude (HA1)" with implicit semantics achieved the poorest result (Macro F1 only 0.19). This is because tweets related to "willingness to seek help" involved in HA1 are expressed indirectly, making it difficult for BERT to accurately capture such semantics [2].

2.2 Detection of Mental Illnesses Based on N-gram

N-gram is a commonly used statistical language modeling tool in natural language processing, referring to a sequence composed of n consecutive basic units (such as words or characters) in a text. Common types include unigrams (n=1, a single word/character), bigrams (n=2, two consecutive units), and trigrams (n=3, three consecutive units). Its core principle is to extract n-gram sequences from text through a sliding window, count the occurrence frequency of different sequences, convert the text into computable vector features, and thereby capture local language patterns and collocation relationships. For example, extracting the bigram "low mood" from tweets to associate with depression-related expressions. In this Twitter-based mental illness detection study, n-grams (such as the combination of unigrams, bigrams, and trigrams with 1-3 words) performed prominently, achieving an AUC of 0.835 in the binary classification task on the Spanish dataset, becoming a key feature for distinguishing between diagnosed and non-diagnosed users. However, it also has limitations such as data sparsity and difficulty in capturing long-distance semantic dependencies, and often needs to be combined with other features to improve effectiveness [5].

In a specific study, 3-gram (a sequence of three consecutive words in text) was used as the sole feature extraction method for detecting depression in social media texts. First, social media comments were preprocessed to remove irrelevant characters and perform text normalization. Then, 3-gram sequences (such as "life is a mess" and "mood is very low") were extracted from the text through a sliding window to capture local contextual associations between words, aiming to mine implicit depression-related semantics. Subsequently, CountVectorizer was used to convert 3-gram sequences into vector features, and SMOTE technology was employed to balance the dataset (expanding minority class samples such as "severe depression" to alleviate the problem of class imbalance). Finally, using the vector features generated by 3-grams as input, a classification model was built to complete the three-classification task of "non-depressed/mildly depressed/severe depressed," realizing text-based depression detection [6].

Relying on the feature extraction capability of 3-grams, the overall average accuracy of the depression detection model reached 56%, ranking 12th in the official competition with an F1 score of 0.401. Among them, the detection effect for the "mildly depressed" category was relatively good, as it could effectively identify depression-related 3-gram patterns in texts of this category. However, its performance for the "non-depressed" category was poor,

ISSN 2959-6157

with low precision and recall. Moreover, since 3-grams can only capture local word order associations, they struggle to mine long-distance semantic dependencies in texts, leaving room for optimization in recognizing depression-related semantics in complex contexts.

2.3 Text Convolutional Neural Network

Convolutional Neural Network (CNN), with its ability to automatically extract features and process high-dimensional data, has become an important tool in mental illness detection. They can mine pathological associations from multimodal data to assist in objective diagnosis. The core of CNNs lies in extracting "low-dimensional details to high-dimensional abstract features" layer by layer through convolutional and pooling layers, eliminating the need for manual feature design. In applications, they can analyze brain imaging (sMRI/fMRI) to identify hippocampal atrophy in schizophrenia patients and abnormal brain networks in depression patients; convert time-series signals such as EEG into images to extract rhythmic abnormalities for anxiety disorder detection; and also enable rapid preliminary screening through facial micro-expressions (such as the drooping corners of the mouth in depression patients) [7].

In the text detection of mental illnesses (taking depression as an example), CNN realizes detection through the process of "automatic text feature extraction - emotion classification - depression association judgment": first, it processes social media texts with emotion labels (such as Twitter tweets), constructs an embedding matrix combined with pre-trained word vectors and inputs it into the model. After the 1D convolutional layer extracts local features and the global max pooling layer screens key features, the fully connected layer outputs emotion classification results. Then, "sadness" emotion is used as the core indicator to assist in judging depression tendency. The model achieved a training accuracy of 98% and a validation accuracy of 91.07% on the SemEval-2018 dataset. For the classification of "sadness" emotion, the precision was 81.7%, the recall was 72.1%, and the F1 score was 76.6%, which can identify implicit depressive expressions. Although there are a small number of missed detections when used alone, when combined with LSTM and other models to form an integrated model, it can improve the detection coverage, providing technical support for early text screening of depression, and its performance is better than traditional machine learning models[8].

2.4 Detection of Mental Illnesses Based on Hierarchical Attention Network

Hierarchical Attention Network (HAN) is a deep learning architecture suitable for mental illness detection. Its core lies in dynamically focusing on key information related to diseases through two levels of attention mechanisms: "bottom-level (word/feature level) and high-level (sentence/sample level)". This solves the problem of noise interference caused by the "equalization processing" of data in traditional models. In detection based on social media texts, it first screens disease-related words such as "insomnia" and "trauma" through word-level attention, and then focuses on texts with high disease relevance (such as posts expressing persistent low mood) through sentence-level attention. In the analysis of EEG signals, it can locate EEG segments related to emotions, reducing the impact of irrelevant fluctuations. Its advantage is that it balances accuracy and interpretability, and weight visualization can clarify the basis for detection (such as key words/EEG segments). It has been validated in detection tasks such as depression and PTSD, and is suitable for multimodal data, providing technical support for early screening of mental illnesses [9].

HAN uses a labeled Twitter dataset of depressed/non-depressed users (finally screening 4208 users, including users' time-series tweets and behavioral characteristics). The tweets are denoised (such as removing stop words and non-ASCII characters), and words are mapped to vectors using GloVe 100-dimensional pre-trained word vectors to construct a three-level data structure of "user-tweet-word". Then, it is divided into word-level encoding and attention, and tweet-level encoding and attention. For word-level encoding and attention, words in a single tweet are processed by BiGRU to capture contextual dependencies (such as the context of depression-related words like "mental illness" and "therapies"). The attention layer calculates the importance weights of words, filtering out meaningless function words such as "and" and "the", and focusing on depressive semantics. For tweet-level encoding and attention, the weighted result of words in a single tweet is used as the tweet vector. BiGRU is applied to multiple tweets of a user to capture temporal correlations (such as the changing trend of depressive emotions). The tweet-level attention layer assigns weights to highlight tweets with high depression signals such as "expressing helplessness and suicidal thoughts". Finally, the text features extracted by HAN can be fused with users' multi-dimensional behavioral features (such as social interaction, emotional expression of emojis, and topic distribution) and input into the sigmoid layer to complete the binary classification of "depressed/non-depressed", while outputting attention weights for result interpretation [10].

On the Twitter depression dataset, HAN, when used alone for text modeling, achieves better classification results than models such as CNN. After fusing with multi-dimensional behavioral features (such as the MDHAN model), the F1 score reaches 89%, which is 10% higher than the traditional MDL model, accurately capturing implicit depressive semantics. Through attention weight visualization (such as heat maps), it can clarify the key classification basis. For example, in the tweets of a depressed user, "One in four experience mental illness" gets the highest tweet weight, and "mental" and "illness" get the highest word weights. At the same time, it can filter out irrelevant tweets (such as daily chats), providing traceable decision-making basis for clinical auxiliary judgment. The final conclusion is that when the number of a user's tweets reaches 200, HAN performs optimally. It can effectively process social media data with "multiple tweet sequences and scattered semantics", avoid interference from irrelevant texts, and is suitable for large-scale community depression screening scenarios [10].

3. Challenges and Solutions

Currently, existing models exhibit poor performance in identifying bipolar disorder and mild depression (F1 < 0.5), primarily due to low annotation consistency (Kappa values ranging from 0.73 to 0.74) and insufficient sample sizes. These models rely heavily on explicit features in text (such as drug names and symptom terms) but lack the ability to learn implicit associations (e.g., psychosocial factors), which may lead to misjudgments—for instance, misclassifying texts related to antidepressants as indicative of depression [2].

Pretraining technologies (e.g., GloVe, BERT embeddings) combined with hierarchical models (HAN) using feature dependency strategies perform best in small-sample, highly imbalanced psychiatric screening tasks, as they balance the recognition performance of majority and minority classes. In contrast, BERT fine-tuning methods are more suitable for large-scale balanced datasets. However, existing approaches still require improvements in modeling label dependencies and identifying minority classes.

To address the challenges in identifying bipolar disorder and mild depression, an enhanced hierarchical model integrating domain knowledge graphs can be constructed. This model first expands high-quality samples through refined annotation guidelines and data augmentation. It then combines a knowledge graph of mental illnesses, symptoms, and psychosocial factors with text features, leveraging an improved hierarchical attention mechanism to capture implicit associations. Additionally, contrastive learning is introduced to strengthen label discriminability, and interpretability tools are incorporated to assist clinical decision-making, thereby balancing the issues of minority class recognition and over-reliance on explicit features.

4. Conclusion

Natural Language Processing holds broad prospects in the field of mental illness detection. In the future, with continuous technological iteration, multimodal fusion technology will integrate diverse data such as text, speech, images, and physiological signals to capture disease characteristics comprehensively, significantly improving detection accuracy. For example, by analyzing changes in intonation and speech rate in speech, combined with emotional expressions in text, mental illnesses can be identified precisely. Meanwhile, large language models, after being fine-tuned with professional psychiatric knowledge, will be able to deeply understand complex medical terminology and patients' vague expressions, and excavate potential symptom clues. The development of edge intelligence and privacy computing will enable data processing on local devices, ensuring patients' privacy and security. In addition, NLP can also contribute to the development of personalized intervention plans, formulating exclusive treatment plans based on patients' language characteristics. It will play a key role in early screening, accurate diagnosis, and long-term treatment monitoring, bringing innovations to the diagnosis and treatment of mental illnesses.

References

[1] Doğu E S, Bektaş A S. The role of sensory processing patterns on mental health in healthy adults. British Journal of Occupational Therapy, 2025, 88(9): 553-562.

[2] Dai H J, Su C H, Lee Y Q, et al. Deep learning-based natural language processing for screening psychiatric patients. Frontiers in Psychiatry, 2021, 11: 533949.

[3] Yang H, Zhu D, He S, et al. Enhancing psychiatric rehabilitation outcomes through a multimodal multitask learning model based on BERT and TabNet: An approach for personalized treatment and improved decision-making. Psychiatry Research, 2024, 336: 115896.

[4] Tseng W H, Lien Y J, Chen C H, et al. A study on the automatic classification of tweets related to mental health literacy ψ. Journal of Educational Media & Library Sciences, 2024, 61(1): 5-27.

[5] Villa-Pérez M E, Trejo L A, Moin M B, et al. Extracting mental health indicators from English and Spanish social media: A machine learning approach. IEEE Access, 2023, 11: 128135-128152.

[6] Malliga S, Shanmugavadivel K, Arunaa S, et al. Team-kec@lt-edi: Detecting signs of depression from social media text, in Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion, 2023: 97-102.

[7] Hassan F, Hussain S F, Qaisar S M. Fusion of multivariate EEG signals for schizophrenia detection using CNN and

ISSN 2959-6157

machine learning techniques. Information Fusion, 2023, 92: 466-478.
[8] Kute R. Mental health analyzer for depression detection based on textual analysis. Journal of Advanced Information Technology, 2022, 13(1).
[9] Ameer I, Arif M, Sidorov G, et al. Mental illness classification on social media texts using deep learning and transfer learning. arXiv preprint arXiv:2207.01012, 2022.
[10] Zogan H, Razzak I, Wang X, et al. Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. World Wide Web, 2022, 25(1): 281-304.