Rising like a Phoenix: Emerging from the Pandemic and Reshaping Human Endeavors with Digital Technologies ICIS 2023

Data Analytics for Business and Societal Challenges

Dec 11th, 12:00 AM

What Symptoms and How Long? An Interpretable AI Approach for Depression Detection in Social Media

Junwei Kuang
Beijing Institute of Technology, kuang_junwei@163.com

Jiaheng Xie
University of Delaware, jxie@udel.edu

Zhijun Yan
Beijing Institute of Technology, yanzhijun@bit.edu.cn

What Symptoms and How Long? An Interpretable AI Approach for Depression Detection in Social Media

Short Paper

Junwei Kuang
Beijing Institute of Technology
Beijing, China
kuang_junwei@163.com

Jiaheng Xie
University of Delaware
DE, USA
jxie@udel.edu

Zhijun Yan
Beijing Institute of Technology
Beijing, China
yanzhijun@bit.edu.cn

Abstract

Depression is the most prevalent and serious mental illness, which induces grave financial and societal ramifications. Depression detection is key for early intervention to mitigate those consequences. Such a high-stake decision inherently necessitates interpretability. Although a few depression detection studies attempt to explain the decision, these explanations misalign with the clinical depression diagnosis criterion that is based on depressive symptoms. To fill this gap, we develop a novel Multi-Scale Temporal Prototype Network (MSTPNet). MSTPNet innovatively detects and interprets depressive symptoms as well as how long they last. Extensive empirical analyses show that MSTPNet outperforms state-of-the-art depression detection methods. This result also reveals new symptoms that are unnoted in the survey approach. We further conduct a user study to demonstrate its superiority over the benchmarks in interpretability. This study contributes to IS literature with a novel interpretable deep learning model for depression detection in social media.

Keywords: social media, depression detection, prototype learning, multi-scale, interpretability

Introduction

Depression is one of the most prevalent mental disorders (WHO, 2017), and brought significant societal and financial consequences. Approximately 280 million people suffer from depression worldwide, accounting for 3.8% of the world's population (Murray, 2022). More than one million people worldwide commit suicide due to depression annually, on par with the number of deaths from cancer (WHO, 2017). The economic toll linked to depression increased from $236.6 billion to $326.2 billion during 2010-2018 in the United States (Greenberg et al., 2021). While many effective depression treatments exist, more than 70% of patients do not seek treatments due to stigmatization (Shen et al., 2017). To mitigate this societal issue and avoid preventable ramifications, depression detection is the key (Picardi et al., 2016).

While surveys remain the primary source of depression detection (Kroenke et al., 2001), social media unleashes the unprecedented potential to expand its reach. Moreover, depressed patients are more willing to communicate on social media compared to offline (Naslund et al., 2016). Many scholars develop depression detection models on social media for early intervention (Chau et al., 2020; Liu et al., 2022). Although achieving satisfying performance, most of these studies rely on black-box methods, which results in limited applicability and potential risk in high-stake scenarios such as healthcare decision-making (Chiong et al., 2021; Zogan et al., 2022). To overcome the non-interpretable dilemma, a few depression detection studies attempt to explain why users are classified as depressed based on the importance score or attention weights of interpretable inputs such as words in a post (Cheng & Chen, 2022). However, existing interpretable models depart from clinical depression diagnosis criteria and receive compromised trust from end users. To tackle their limitations, there has recently been a rising interest in utilizing symptoms for interpreting depression detection. Pioneering studies have shown the potential benefits of improving accuracy, generalizability, and interpretability (Nguyen et al., 2022; Zhang et al., 2022b). Therefore, our research objective is to develop an ***interpretable depression detection model in social media based on symptom-based depression diagnostic criteria***.

The symptom-based interpretable methods for depression detection can be categorized into dictionary-based, similarity-based, and classification-based (Shen et al., 2017). The core of these methods is to identify depressive symptoms from user-generated posts on social media. However, these methods still face three limitations. First, prior methods only identify pre-defined symptoms. However, depressive symptoms may evolve over time. Second, previous methods rely on domain-specific knowledge, which require significant labor costs and suffer from poor generalizability. Third, extant methods focus on *what symptoms* users present while neglecting *how long* these symptoms last (Kroenke et al., 2001). Fortunately, user-generated posts on social media can reveal such "how long" aspects of depressive symptoms. As shown in Figure 1, the user reported the disturbed sleep symptom numerous times, ranging from Feb 14 to May 14. Certain periods (e.g., May 12 to May 14) show denser symptom mentions than others.

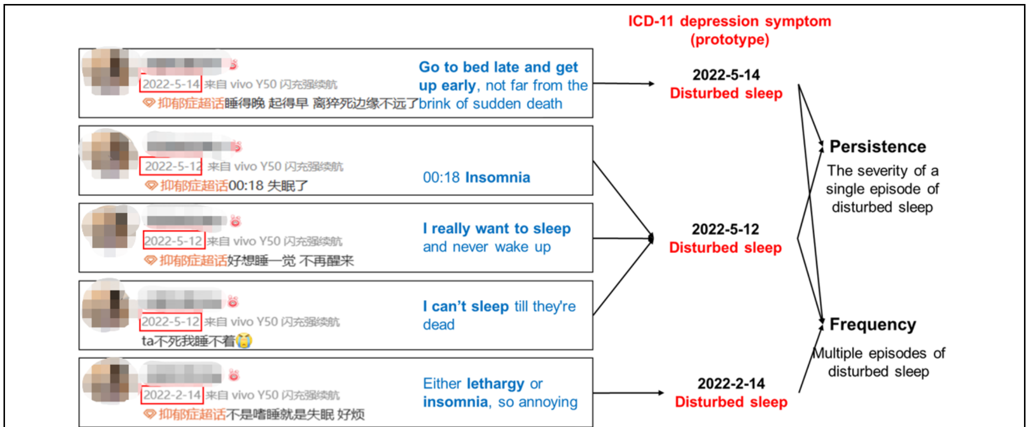

Figure 1. Typical Posts Disclosing Depressive Symptoms in Social Media

The abovementioned limitations motivate us to develop a novel interpretable depression detection method. Following the computational design science paradigm and prior IS research on health analytics (Yu et al., 2023), we propose and rigorously evaluate a novel interpretable model, Multi-Scale Temporal Prototype Network (MSTPNet). MSTPNet is built upon an emergent stream of case-based interpretable models, prototype learning (Ming et al., 2019), which interprets the prediction for new inputs by comparing them with a few learned prototypes. In this study, typical posts disclosing depressive symptoms can be recognized as prototypes. To consider how long the symptoms last, MSTPNet modifies standard prototype learning methods by devising two novel layers: a temporal segmentation layer that eliminates the negative effects of irrelevant and redundant posts on symptom identification to facilitate period-level analysis (i.e., "What symptoms did the user suffer in a period?"), and a multi-scale temporal prototype layer that captures the temporal distribution of symptoms. In practice, our method can be implemented in social media to detect depressed patients and interpret their temporal symptoms. When implementing intervention, platform managers need to combine human intelligence to judge rather than rely entirely on artificial intelligence.

Literature Review

Social media-based depression detection is broadly classified into traditional machine learning, black-box deep learning, and interpretable deep learning. The traditional machine learning-based depression detection model mostly relies on effective input features (Li et al., 2019). Table 1 summarizes the traditional machine learning-based method for depression detection in social media.

Reference	Dataset	Sample (depression/non)	Input features	Methods
Choudhury et al. (2013)	Twitter	476 (171/305)	Emotion, Depression language, Language style	SVM
Tsugawa et al. (2015)	Twitter	209 (81/128)	Emotions, Linguistic style, Topic, social Network	LDA, SVM
Chen et al. (2018)	Twitter	1200 (600/600)	Emotion swings, LIWC	SVM, RF
Chau et al. (2020)	Blog	804 (274/530)	N-gram, Lexicon based, LIWC	SVM, Rule-based, GA
Chiong et al. (2021)	Twitter	2804 (1402/1402)	N-gram	SVM, DT, NB, KNN,

Table 1. Traditional Machine Learning Methods in Depression Detection

However, these studies have shown unsatisfactory predictive power mainly because hand-crafted features and traditional machine learning models are not complex enough to capture high-level interactions between features. Black-Box deep learning methods have demonstrated significantly higher predictive power in depression detection (Malhotra & Jindal, 2022). These improvements have benefited from the development of embedding techniques and the utilization of various neural network architectures. Table 2 summarizes recent black-box deep learning-based depression detection methods in social media.

Reference	Dataset	Sample (depression/non)	Input features	Methods
Orabi et al. (2018)	Twitter	899 (327/572)	Text	CNN/RNN
Chiu et al. (2021)	Instagram	520 (260/260)	Text, Image, Posting time	LSTM with temporal weighting
Ghosh and Anwar (2021)	Twitter	6562 (1402/5160)	Text	LSTM
Zogan et al. (2022)	Twitter	4800 (2500/2300)	Text, Image	HAN
Kour and Gupta (2022)	Twitter	1681 (941/740)	Text	CNN + Bi-LSTM

Table 2. Black-Box Deep Learning Methods in Depression Detection

Despite their satisfying performance, their lack of interpretability limits their applicability in high-stake decision-making scenarios (Rudin, 2019). Interpretable deep learning methods refer to deep learning methods that provide a certain explanation (Li et al., 2022). Table 3 summarizes and contrasts recent interpretable deep learning-based methods and our study in social media-based depression detection.

Reference	**Type**	**Method**	**Usage**	**Explanations**
Adarsh et al. (2023)	Approximation	LIME	Post-hoc	Important raw inputs
Cheng and Chen (2022)	Attention	Attention	Intrinsic	Important raw inputs
Zogan et al. (2022)	Attention	HAN	Intrinsic	Important raw inputs
Shen et al. (2017)	Symptom	Dictionary-based	Post-hoc	Predicted symptoms
Zhang et al. (2022b)	Symptom	Classification-based	Post-hoc	Predicted symptoms
Zhang et al. (2022a)	Symptom	Similarity-based	Post-hoc	Predicted symptoms
Our study	**Symptom**	**Similarity-based**	**Intrinsic**	**More symptoms, and how long**

Table 3. Interpretable Deep Learning Methods in Depression Detection

Symptom-based interpretable deep learning methods align with clinical depression criteria, but still face two limitations. First, they generally require high labor costs and only identify pre-defined symptoms, neglecting new symptoms unnoted in offline depression screening questionnaires in the online setting. Second, symptom-based interpretable methods focus only on the type of depressive symptoms users suffer, neglecting how long these symptoms last, which is equally critical for a clinical depression diagnosis. These limitations motivate us to develop a novel interpretable depression detection method that is capable of discovering depressive symptoms in a data-driven manner while capturing how long these symptoms last.

We resort to an emergent interpretable model paradigm that is closely related to our task: prototype learning. Prototype learning methods learn prototypes that have clear semantic meanings, and intrinsic explanations are generated based on the comparison between input and each prototype (Nauta et al., 2021). Chen et al. (2019) originally propose ProtoPNet, which explains the contribution of prototypical parts of the predicted image by comparing the learned prototypes. Multiple prototype learning variants have also been proposed for various tasks. Typical posts disclosing depressive symptoms can be recognized as prototypes in our study. By calculating how similar a user's posts are to these prototypes, this user's depressive symptoms can be inferred, which serves as a natural interpretation mechanism. Table 4 contrasts major prototype learning methods with our method.

Reference	Method	Novelty	Input	TD*
Chen et al. (2019)	ProtoPNet	Prototype for image classification	An image	No
Hase et al. (2019)	HPNet	Hierarchical prototype	An image	No
Ming et al. (2019)	ProSeNet	Prototype for text classification	A piece of text	No
Zhang et al. (2020)	TapNet	Attentional prototype	A time series of ECG	No
Nauta et al. (2021)	ProtoTree	Prototype and decision tree	An image	No
Trinh et al. (2021)	DPNet	Dynamic prototype	A clip of a video	No
Deng et al. (2022)	K-HPN	Pairwise prototype	A piece of text	No
Our study	**MSTPNet**	**Multi-scale temporal prototypes**	**A sequence of text documents**	**Yes**

Table 4. Existing Prototype Learning Methods vs. Our Method

* TD stands for "Temporal Distribution", indicating whether a model considers the temporal distribution of the prototype, which includes frequency and persistence of appearance at the period level

The majority of prototype learning methods focus on static subjects, such as an image and a piece of text. When applied to our study, these methods only consider whether depressive symptoms appear, neglecting how long each symptom last (Chen et al., 2019). While a few prototype learning methods process dynamic subjects such as video, these methods focus on directly identifying complex prototypes with temporal properties, rather than analyzing the temporal distribution of prototypes after identifying them. Our method aims to incorporate the temporal distribution of symptoms into the prototype learning method to effectively capture how long depressive symptoms last to improve the predictive power and interpretability.

The MSTPNet Approach

Figure 2 shows the architecture of MSTPNet, which features four building blocks. The feature learning layer aims to represent each post as an embedding vector with a fixed length and rich semantic meaning. Different from analyzing each post independently, our proposed temporal segmentation layer assigns posts into different periods based on the semantic similarity and time interval between posts, which facilitate period-level analysis. Instead of learning complex dynamic prototypes (e.g., "long-term disturbed sleep") directly, our proposed multi-scale temporal prototype layer breaks the task down into two parts. We first infer depressive symptoms (e.g., "disturbed sleep") in each period by comparing posts with learned prototypes, and then explicitly measure the frequency (e.g., the proportion of periods where disturbed sleep appears) and persistence (e.g., the number of continuous periods where disturbed sleep all appears) of each symptom. Based on the above interpretable temporal measurement of each symptom, the classification layer classifies a user into depression or non-depression categories.

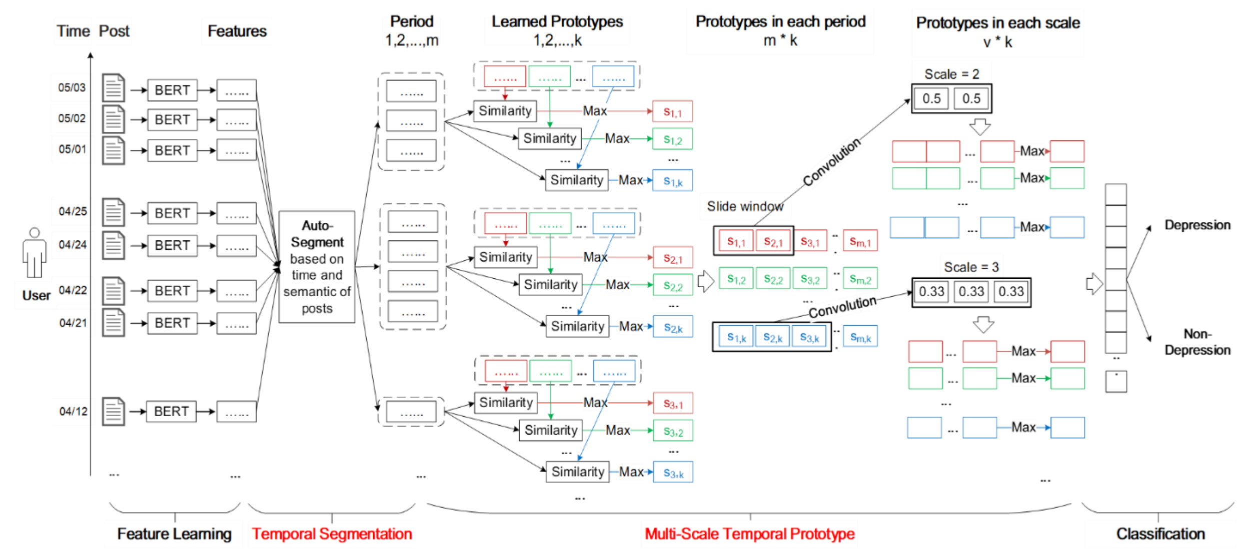

Figure 2. MSTPNet Architecture

To learn an effective representation for each post, we deploy a feature learning layer using the cutting-edge pre-trained language model BERT (Devlin et al., 2019). Specifically, for a post X_i:

$$H_i = BERT(X_i) \tag{1}$$

Our temporal segmentation layer builds upon a bottom-up hierarchical clustering algorithm (Shetty & Singh, 2021) to segment the social media posts $u = (H_1, t_1; H_2, t_2; \dots; H_n, t_n)$ into m periods $u = (C_1, C_2, \dots, C_m)$, where $C_i = (H_{i,1}, t_{i,1}; H_{i,2}, t_{i,2}; \dots; H_{i,l}, t_{i,l})$. The key to segmentation methods is the distance measurement between different posts. We propose a new measurement that combines both semantic similarity and the time interval between posts in Formula (2), (3), and (4).

$$sim_{sem}^{i,j} = \frac{H_i * H_j}{\|H_i\| * \|H_j\|} \tag{2}$$

$$sim_{time}^{i,j} = \exp\left(-\frac{|t_i - t_j|}{w_d}\right) \tag{3}$$

$$sim^{i,j} = w_a * sim_{time}^{i,j} + (1 - w_a) * sim_{sem}^{i,j} \tag{4}$$

In each iteration, the method calculates the similarity between each pair of segments, and then merges the most similar pair into a new segment, until the time distance between the two segments exceeds the pre-defined length h of periods. The remaining clusters $(C_1, C_2, \dots, C_m)$ are the segmentation results, where C_i is the i-th segment of the focal user, and $X_{i,j}$ is the j-th post in the segment C_i.

Then, we define k prototypes $P = (p_1, p_2, \dots, p_k)$ to be leaned, where each prototype is learnable parameters with the same length as the latent representation of each post. We can assign p_i with the closest post in the training data to translate prototypes and make them interpretable (Trinh et al., 2021). Next, based on the learned prototypes, we infer the existence $s_{m,k}$ of the symptom k in period m that contains l posts, as shown in Formulas (5). The $s_{m,j,k}$ denotes the similarity between depressive symptom prototype p_k and latent representation $H_{m,j}$ of the j-th post in m-th period C_m by using L2 distance (Ming et al., 2019).

$$s_{m,k} = \max_{j=1,2,\dots,l} exp\left(-\left\|H_{m,j} - p_k\right\|^2\right) \tag{5}$$

Beyond one period or all periods, we focus on a specified number of consecutive periods to measure the frequency and persistence of depressive symptoms. The number of consecutive periods is called “scale” in this study. Different scales enable analysis at different granularity and can capture comprehensive clues to detect depression. Therefore, we employ multiple scales (i.e., multi-scale) with different sizes for period-level analysis, which is conceptually similar to the filters with different sizes in CNN to analyze image data.

The difference is that our "filters" are not learnable parameters but are explicitly set to get the average value over continuous periods, which is easy for humans to understand. Specifically, let W be the set of scales used in our model, and w_j denotes the size of the j-th scale. We calculate the existence (i.e., average similarity) of depressive symptoms in each pair with a window length of the scale, and then take the highest value as the existence (i.e., $g_{j,k}$) of the depressive symptom k on the scale w_j, as shown in Formula (6).

$$g_{j,k} = \max_{m=1,2,\dots,M-w_j+1} \frac{1}{w_j} \sum_{m}^{m+w_j-1} s_{m,k} \tag{6}$$

We let $G = (g_{1,1}, g_{1,2}, \dots g_{1,K}, \dots g_{j,K,\dots,}g_{J,K})$, where J is the number of scales. The classification layer computes the probability of depression given all $g_{j,k}$ (G) of a focal user, as shown in formulas (7) and (8):

$$\hat{y}_i = \frac{\exp(z_i)}{\sum_{s=0}^{1} \exp(z_s)} \tag{7}$$

$$Z = QG \tag{8}$$

Following Ming et al. (2019), the loss function of MSTPNet to be minimized is defined based on the binary cross-entropy (CE) loss with four additional regularization terms. Specifically,

$$Loss = CE + \lambda_c R_c + \lambda_e R_e + \lambda_d R_d + \lambda_{l_1} \|Q\|_1 \tag{9}$$

$$CE = \sum_{(X,y)\in D} y \log \hat{y} + (1-y)\log(1-\hat{y}) \tag{10}$$

where λ_c, λ_e, λ_d and λ_{l_1} are hyperparameters that determine the weight of the regularizations.

Empirical Analysis

We use the WU3D, an annotated dataset regarding depression detection in a Chinese social media platform (Wang et al., 2022). It contains chronological sequences of posts from 10,325 depressed users and a random control group of 22,245 users. We set imbalance ratios as 1:8, which approximates the ratio of adults with depression risk in China (Fu et al., 2023).We split this dataset into 60% for training, 20% for validation, and 20% for test. We set K=70, h=15, w_a=0.4, and the scales contain 1, 2, 3, 5, 8, 12, 16 and 20. The evaluation results are reported in Tables 5. MSTPNet outperforms benchmark models in F1 score and accuracy and outperforms interpretable deep learning models in all metrics.

Models	F1	Precision	Recall	Accuracy
Yang et al. (2020)	0.412***	0.887***	0.268***	0.918***
Chen et al. (2018)	0.508***	0.891***	0.355***	0.926***
Chiong et al. (2021)	0.380***	0.363***	0.399***	0.861***
Chau et al. (2020)	0.706***	0.604***	0.850	0.924***
Orabi et al. (2018)	0.828***	0.878***	0.785	0.965**
Chiu et al. (2021)	0.822**	0.936*	0.732**	0.966**
Ghosh and Anwar (2021)	0.820**	0.921**	0.741***	0.965***
Naseem et al. (2022)	0.826**	0.963	0.723**	0.967*
Cheng and Chen (2022)	0.806***	0.910**	0.723**	0.963**
Zogan et al. (2022)	0.795***	0.840***	0.754*	0.958**
Ming et al. (2019)	0.816**	0.929*	0.729***	0.965**
Chen et al. (2019)	0.735***	0.876***	0.633***	0.951**
Trinh et al. (2021)	0.675***	0.774***	0.598***	0.938***
MSTPNet	**0.851**	**0.957**	**0.766**	**0.971**

Table 5. SOTA Methods vs. Our Method

Note: *p < 0.05; **p < 0.01; ***p < 0.001

We further perform ablation studies to show their effectiveness as shown in Table 6. We remove the temporal segmentation layer to validate the effectiveness of period-level analysis. We also replace the multi-scale temporal prototype (MS) layer with a common prototype learning layer. We test two options: the maximum or mean existence strength of prototype.

Models	F1	Precision	Recall	Accuracy
MSTPNet (Ours)	**0.851**	**0.957**	**0.766**	**0.971**
MSTPNet removing temporal segmentation layer	0.801***	0.923***	0.702***	0.962***
MSTPNet removing MS using Max	0.760***	0.868***	0.676***	0.954***
MSTPNet removing MS using Mean	0.690***	0.846***	0.583***	0.944***

Table 6. Ablation Studies

Note: *p < 0.05; **p < 0.01; ***p < 0.001

Our MSTPNet provides a level of interpretability that is absent in other interpretable deep models. Figure 3 provides a visual comparison of different types of interpretation. Our interpretation is capable of capturing what symptoms users suffer and how long these symptoms last, which aligns with the clinical depression diagnosis criterion. Moreover, our MSTPNet is based on prototype learning and can show new depressive symptoms rather than just pre-defined symptoms.

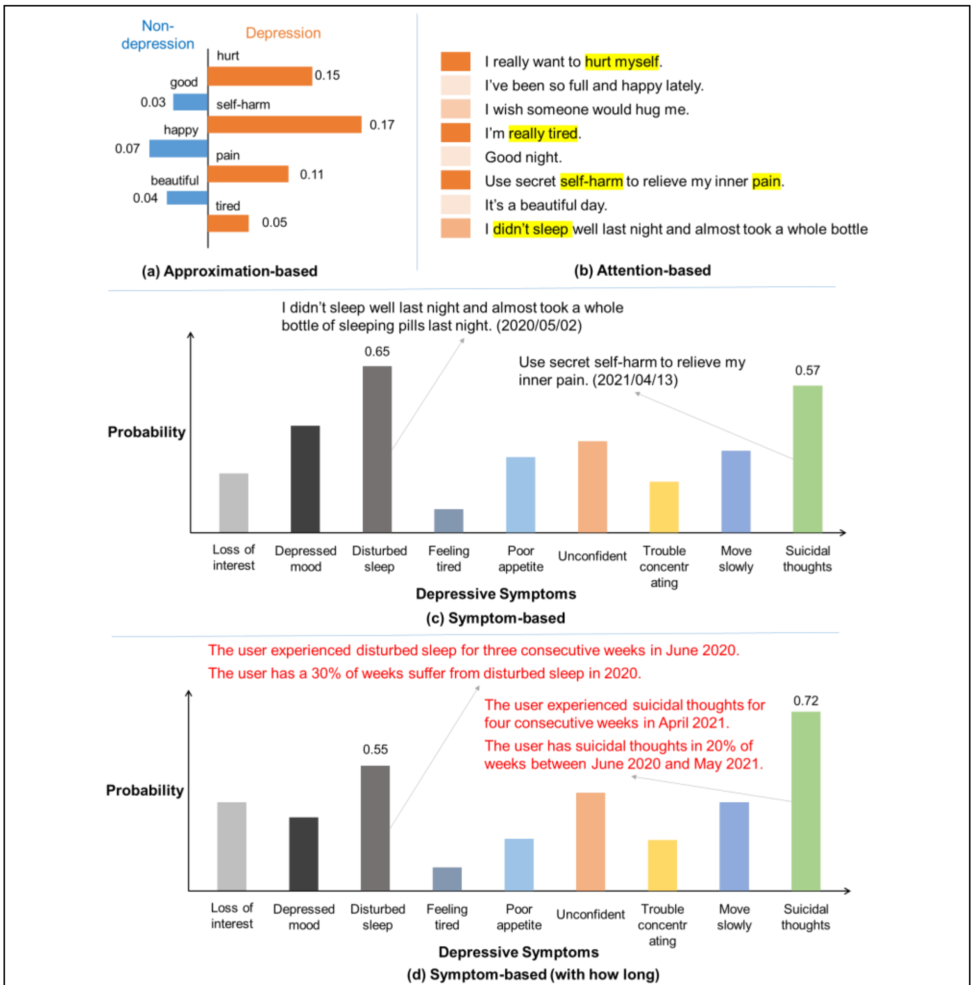

Figure 3. MSTPNet Architecture

Discussion and Conclusion

We propose a novel interpretable deep learning method to detect and interpret depression based on what symptoms the user has and how long these related symptoms last. We conduct extensive evaluations to demonstrate the superior predictive power and interpretability of our method over state-of-the-art benchmarks. Our study establishes a few generalized design principles: (1) A temporal segmentation module could facilitate period-level analysis and mitigate the effect of redundant and irrelevant information; (2) It's cost-effective and flexible to explicitly separate a complex task into two related simple tasks; (3) Showing the temporal distribution of prototypes could improve interpretability and boost the trust and perceived helpfulness. These design principles prescribe how to predict and interpret the hidden state of a user from a sequence of user-related data, such as social media posts, electric health records.

Acknowledgements

This study was supported by National Natural Science Foundation of China (No: 72110107003, 71872013).

References

Adarsh, V., Arun Kumar, P., Lavanya, V., et al. (2023). Fair and explainable depression detection in social media. *Information Processing & Management*, 60(1), 103168.

Chau, M., Li, T. M. H., Wong, P. W. C., et al. (2020). Finding people with emotional distress in online social media: A design combining machine learning and rule-based classification. *MIS Quarterly*, 44(2), 933-956.

Chen, C., Li, O., Tao, D., et al. (2019) This looks like that: Deep learning for interpretable image recognition. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 8930-8941.

Chen, X., Sykora, M., Jackson, T. W., et al. (2018) What about mood swings: Identifying depression on twitter with temporal measures of emotions. *Proceedings of the The Web Conference 2018*, 1653-1660.

Cheng, J. C., & Chen, A. L. (2022). Multimodal time-aware attention networks for depression detection. *Journal of Intelligent Information Systems*, 59(2), 319-339.

Chiong, R., Budhi, G. S., Dhakal, S., et al. (2021). A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. *Computers in Biology and Medicine*, 135(8), 104499.

Chiu, C. Y., Lane, H. Y., Koh, J. L., et al. (2021). Multimodal depression detection on instagram considering time interval of posts. *Journal of Intelligent Information Systems*, 56, 25-47.

Choudhury, M. D., Counts, S., & Horvitz, E. (2013) Social media as a measurement tool of depression in populations. *Proceedings of the 5th Annual ACM Web Science Conference*, Paris, France, 47–56.

Deng, S., Zhang, N., Chen, H., et al. (2022). Low-resource extraction with knowledge-aware pairwise prototype learning. *Knowledge-Based Systems*, 235, 107584.

Devlin, J., Chang, M.-W., Lee, K., et al. (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171–4186.

Fu, X., Zhang, K., Zhang, X., et al. (2023). *Report on national mental health development in china (2021-2022)*. S. S. A. Press.

Ghosh, S., & Anwar, T. (2021). Depression intensity estimation via social media: A deep learning approach. *IEEE Transactions on Computational Social Systems*, 8(6), 1465-1474.

Greenberg, P. E., Fournier, A.-A., Sisitsky, T., et al. (2021). The economic burden of adults with major depressive disorder in the united states (2010 and 2018). *Pharmaco Economics*, 39(6), 653-665.

Hase, P., Chen, C., Li, O., et al. (2019) Interpretable image recognition with hierarchical prototypes. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 32-40.

Kour, H., & Gupta, M. K. (2022). An hybrid deep learning approach for depression prediction from user tweets using feature-rich cnn and bi-directional lstm. *Multimedia Tools and Applications*, 81(17), 23649-23685.

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The phq-9 - validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9), 606-613.

Li, X., Xiong, H., Li, X., et al. (2022). Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems*, 64(12), 3197-3234.

Li, X. W., Zhang, X., Zhu, J., et al. (2019). Depression recognition using machine learning methods with different feature generation strategies. *Artificial Intelligence in Medicine*, 99, 101696.

Liu, D. X., Feng, X. L., Ahmed, F., et al. (2022). Detecting and measuring depression on social media using a machine learning approach: Systematic review. *Jmir Mental Health*, 9(3), e27244.

Malhotra, A., & Jindal, R. (2022). Deep learning techniques for suicide and depression detection from online social media: A scoping review. *Applied Soft Computing*, 130, 109713.

Ming, Y., Xu, P., Qu, H., et al. (2019) Interpretable and steerable sequence learning via prototypes. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 903-913.

Murray, C. J. L. (2022). The global burden of disease study at 30 years. *Nature Medicine*, 28(10), 2019-2026.

Naseem, U., Dunn, A. G., Kim, J., et al. (2022) Early identification of depression severity levels on reddit using ordinal classification. *Proceedings of the ACM Web Conference 2022*, 2563-2572.

Naslund, J. A., Aschbrenner, K. A., Marsch, L. A., et al. (2016). The future of mental health care: Peer-to-peer support and social media. *Epidemiology and psychiatric sciences*, 25(2), 113-122.

Nauta, M., van Bree, R., & Seifert, C. (2021) Neural prototype trees for interpretable fine-grained image recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14933-14943.

Nguyen, T., Yates, A., Zirikly, A., et al. (2022) Improving the generalizability of depression detection by leveraging clinical questionnaires. *60th Annual Meeting of the Association for Computational Linguistic*, 8446-8459.

Orabi, A. H., Buddhitha, P., Orabi, M. H., et al. (2018) Deep learning for depression detection of twitter users. *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 88-97.

Picardi, A., Lega, I., Tarsitani, L., et al. (2016). A randomised controlled trial of the effectiveness of a program for early detection and treatment of depression in primary care. *Journal of Affective Disorders*, 198, 96-101.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206-215.

Shen, G., Jia, J., Nie, L., et al. (2017) Depression detection via harvesting social media: A multimodal dictionary learning solution. *IJCAI*, 3838-3844.

Shetty, P., & Singh, S. (2021). Hierarchical clustering: A survey. *International Journal of Applied Research*, 7(4), 178-181.

Trinh, L., Tsang, M., Rambhatla, S., et al. (2021) Interpretable and trustworthy deepfake detection via dynamic prototypes. *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1973-1983.

Tsugawa, S., Kikuchi, Y., Kishino, F., et al. (2015) Recognizing depression from twitter activity. *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 3187-3196.

Wang, Y. D., Wang, Z. Y., Li, C. H., et al. (2022). Online social network individual depression detection using a multitask heterogenous modality fusion approach. *Information Sciences*, 609, 727-749.

WHO. (2017). *Depression and other common mental disorders: Global health estimates*. WHO.

Yang, X., McEwen, R., Ong, L. R., et al. (2020). A big data analytics framework for detecting user-level depression from social networks. *International Journal of Information Management*, 54, 102141.

Yu, S., Chai, Y., Samtani, S., et al. (2023). Motion sensor-based fall prevention for senior care: A hidden markov model with generative adversarial network (hmm-gan) approach. *Information Systems Research*, Online.

Zhang, X., Gao, Y., Lin, J., et al. (2020) Tapnet: Multivariate time series classification with attentional prototypical network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 6845-6852.

Zhang, Z., Chen, s., Mengyue Wu, et al. (2022a) Psychiatric scale guided risky post screening for early detection of depression. *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 5220-5226.

Zhang, Z., Chen, S., Wu, M., et al. (2022b) Symptom identification for interpretable detection of multiple mental disorders on social media. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 9970-9985.

Zogan, H., Razzak, I., Wang, X. Z., et al. (2022). Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. *World Wide Web-Internet and Web Information Systems*, 25(1), 281-304.