Explainability in NLP for Mental Health: A Comprehensive Survey

Prajval Bolegave, Pushpak Bhattacharya
Computer Science and Engineering, IIT Bombay, India
{prajval, pb}@cse.iitb.ac.in

Abstract

The intersection of Natural Language Processing (NLP) and mental health has emerged as a critical domain for developing AI-powered therapeutic interventions, diagnostic tools, and supportive systems. However, the "black box" nature of many state-of-the-art NLP models poses significant challenges in clinical settings where transparency, trust, and accountability are paramount. This survey provides a comprehensive overview of explainability techniques in NLP applications for mental health, examining both model-agnostic and model-specific interpretability methods. We analyze current approaches including LIME, SHAP, influence functions, gradient-based methods, and Layer-wise Relevance Propagation (LRP) in the context of mental health applications such as depression detection, suicide risk assessment, empathetic response generation, and therapeutic dialogue systems. Through systematic analysis of recent literature, we identify key challenges, opportunities, and future directions for developing more interpretable and trustworthy NLP systems in mental healthcare.

1 Introduction

Mental health disorders affect millions of people worldwide, with the World Health Organization estimating that one in four people will be affected by mental disorders at some point in their lives. The advent of social media and digital communication platforms has created unprecedented opportunities for early detection, monitoring, and intervention in mental health care through computational approaches (Chancellor and De Choudhury, 2020). Natural Language Processing has emerged as a powerful tool for analyzing textual data from various sources including social media posts, clinical notes, therapy transcripts, and conversational interfaces to understand and support mental health conditions.

Recent advances in deep learning and large language models have shown remarkable performance in various mental health NLP tasks, from detecting depression and anxiety in social media posts (Shen et al., 2017) to generating empathetic responses in therapeutic chatbots (Rashkin et al., 2019). However, the increasing complexity of these models has led to a significant interpretability gap, making it difficult for healthcare professionals to understand and trust the model's decisions.

The need for explainable AI in healthcare is particularly acute due to several factors: (1) the high-stakes nature of mental health decisions, (2) regulatory requirements for transparency in medical AI systems, (3) the need for healthcare providers to understand and validate AI recommendations, and (4) the importance of building trust between patients and AI-powered therapeutic tools (Holzinger et al., 2017).

This survey aims to bridge the gap between explainable AI techniques and their applications in mental health NLP. We provide a comprehensive review of current explainability methods, analyze their effectiveness in mental health contexts, and identify key challenges and future research directions.

2 Background and Related Work

2.1 Mental Health Applications in NLP

The application of NLP in mental health spans several key areas:

Depression and Anxiety Detection: Early work focused on identifying linguistic markers of depression and anxiety in social media text. Studies have shown that individuals with depression tend to use more first-person pronouns, negative emotion words, and absolute language (Coppersmith et al., 2014).

Suicide Risk Assessment: NLP models have been developed to identify individuals at risk of suicide by analyzing their social media posts, with systems achieving reasonable accuracy but requiring careful consideration of false positives and negatives (Ji et al., 2021).

Empathetic and Therapeutic Dialogue Systems: Recent research has focused on developing conversational AI systems that can provide empathetic responses and therapeutic support. Work by (Sharma et al., 2022) on motivational virtual assistants and (Majumder et al., 2020) on empathetic response generation has shown promising results.

Clinical Decision Support: NLP systems have been developed to assist clinicians in diagnosis and treatment planning by analyzing clinical notes and patient communications (Coppersmith et al., 2018).

2.2 The Need for Explainability in Mental Health NLP

The deployment of NLP systems in mental health contexts raises several critical concerns that necessitate explainable AI approaches:

Clinical Trust and Adoption: Healthcare professionals require understanding of how AI systems reach their conclusions to make informed decisions about patient care (Tonekaboni et al., 2019).

Bias and Fairness: Mental health NLP systems may inadvertently perpetuate biases related to demographics, cultural background, or linguistic patterns, making bias detection and mitigation crucial (Benton et al., 2019).

Patient Safety: Incorrect predictions in mental health applications can have severe consequences, making it essential to understand when and why models fail (Luo et al., 2016).

Regulatory Compliance: Increasing regulatory requirements for AI transparency in healthcare demand explainable systems (Watson et al., 2019).

3 Explainability Methods in NLP

This section provides a comprehensive overview of key explainability techniques that have been applied or show promise for mental health NLP applications. We present both the mathematical foundations and practical applications of these methods in the context of mental health text analysis.

3.1 Model-Agnostic Methods

Model-agnostic methods can be applied to any machine learning model regardless of its internal architecture, making them particularly valuable for explaining complex deep learning models used in mental health NLP.

3.1.1 LIME (Local Interpretable Model-agnostic Explanations)

LIME (Ribeiro et al., 2016) generates explanations by learning local interpretable models around individual predictions. For a given input text x and model f, LIME aims to find an explanation $g \in G$ that minimizes:

$$\xi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{1}$$

where $L(f, g, \pi_x)$ is the locality-aware loss function, π_x is a proximity measure, G is the class of interpretable models, and $\Omega(g)$ is a complexity measure.

For NLP applications, LIME generates perturbed samples z' by randomly removing words from the original text x. The proximity measure is typically defined as:

$$\pi_x(z) = \exp\left(-\frac{D(x, z)^2}{\sigma^2}\right) \tag{2}$$

where $D(x, z)$ represents the distance between the original text and the perturbed version, often measured as the number of differing words.

The interpretable model g is typically a linear model of the form:

$$g(z') = w_0 + \sum_{i=1}^{d'} w_i z'_i \tag{3}$$

where $z'_i \in \{0, 1\}$ indicates the presence or absence of the i-th interpretable feature (word), and w_i represents the contribution of that feature to the prediction.

In mental health NLP, LIME has been particularly effective for analyzing depression detection models.

3.1.2 SHAP (SHapley Additive exPlanations)

SHAP (Lundberg and Lee, 2017) provides a unified framework for feature importance based on cooperative game theory. The SHAP value for feature i is defined as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \tag{4}$$

where N is the set of all features, S is a subset of features not including feature i, and $f(S)$ is the model's prediction when only features in set S are present.

SHAP values satisfy four desirable axioms:

- **Efficiency:** $\sum_{i=1}^{N} \phi_i = f(N) - f(\emptyset)$
- **Symmetry:** If $f(S \cup \{i\}) = f(S \cup \{j\})$ for all $S \subseteq N \setminus \{i, j\}$, then $\phi_i = \phi_j$
- **Dummy:** If $f(S \cup \{i\}) = f(S)$ for all $S \subseteq N \setminus \{i\}$, then $\phi_i = 0$
- **Additivity:** For $f = f_1 + f_2$, we have $\phi_i[f] = \phi_i[f_1] + \phi_i[f_2]$

For text data, computing exact SHAP values is computationally intractable due to the exponential number of possible feature subsets. Therefore, approximation methods are used:

KernelSHAP: Uses weighted linear regression to approximate SHAP values:

$$\phi_i \approx \arg\min_{\phi} \sum_{z \in Z} \pi(z) \| f(z) - \phi_0 - \sum_{j=1}^{M} \phi_j z_j \|^2 \tag{5}$$

where Z is a set of perturbed samples, $\pi(z)$ is the kernel weight, and M is the number of simplified input features.

TreeSHAP: For tree-based models, provides an exact polynomial-time algorithm by computing:

$$\phi_i = \sum_{l \in L} \frac{p_l(S_l^i \cup \{i\}) - p_l(S_l^i)}{|S_l^i|!(M - |S_l^i| - 1)!/M!} \tag{6}$$

where L is the set of leaf nodes, p_l is the contribution of leaf l, and S_l^i is the set of features on the path to leaf l that come before feature i.

In mental health applications, SHAP has been used to analyze suicide risk prediction models.

3.1.3 Influence Functions

Influence functions (Koh and Liang, 2017) measure the effect of training examples on model predictions. For a model with parameters $\hat{\theta}$ that minimize the empirical risk $\frac{1}{n}\sum_{i=1}^{n} L(z_i, \theta)$, the influence of training point z on the prediction at test point z_{test} is:

$$\mathcal{I}_{up,params}(z, z_{test}) = -\nabla_\theta L(z_{test}, \hat{\theta})^T H_{\hat{\theta}}^{-1} \nabla_\theta L(z, \hat{\theta}) \tag{7}$$

where $H_{\hat{\theta}} = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta^2 L(z_i, \hat{\theta})$ is the Hessian matrix.

For computational efficiency, the Hessian inverse is approximated using:

$$H_{\hat{\theta}}^{-1} v \approx \sum_{j=0}^{J} (I - \alpha H_{\hat{\theta}})^j v \tag{8}$$

where α is a damping factor and J is the number of iterations.

In mental health NLP, influence functions have been used to identify potentially biased or mislabeled training examples. For instance, in a depression detection model, influence function analysis revealed that training examples containing demographic information (e.g., "I'm a 45-year-old woman feeling sad") had disproportionately high influence scores, suggesting the model was learning demographic shortcuts rather than clinical symptoms.

3.2 Gradient-Based Methods

Gradient-based methods leverage the gradient information from neural networks to understand feature importance and model behavior.

3.2.1 Gradient-based Attribution

The simplest gradient-based method computes feature importance as:

$$R_i^{(simple)} = \frac{\partial f(x)}{\partial x_i} \tag{9}$$

However, this approach suffers from the saturation problem where gradients become very small in saturated regions of activation functions.

3.2.2 Gradient × Input

To address saturation issues, the gradient is multiplied by the input:

$$R_i^{(grad \times input)} = x_i \cdot \frac{\partial f(x)}{\partial x_i} \tag{10}$$

This method provides better attributions but can still be noisy for complex models.

3.2.3 Integrated Gradients

Integrated Gradients (Sundararajan et al., 2017) addresses the limitations of vanilla gradients by integrating gradients along a path from a baseline input x' to the actual input x:

$$IG_i(x) = (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial f(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha \tag{11}$$

In practice, this integral is approximated using the trapezoidal rule:

$$IG_i(x) \approx (x_i - x'_i) \times \sum_{k=1}^{m} \frac{1}{m} \times \frac{\partial f\left(x' + \frac{k}{m} \times (x - x')\right)}{\partial x_i} \tag{12}$$

where m is the number of steps in the approximation.

For text data, the baseline x' is typically chosen as the zero vector (representing the absence of all words) or a neutral text. Integrated Gradients satisfies two important axioms:

- **Sensitivity:** If x and x' differ in one feature and have different predictions, then that feature has non-zero attribution
- **Implementation Invariance:** Attributions are identical for functionally equivalent networks

In mental health dialogue systems, Integrated Gradients has been used to understand which words in user messages trigger empathetic responses. Analysis of a therapeutic chatbot revealed that words like "struggling" and "overwhelmed" had high integrated gradient scores for empathy generation.

3.2.4 Layer-wise Relevance Propagation (LRP)

LRP (Bach et al., 2015) decomposes the model's prediction by propagating relevance scores backward through the network layers. The basic principle is that relevance is conserved: $\sum_i R_i^{(l)} = \sum_j R_j^{(l+1)}$ for consecutive layers l and $l+1$.

For a neuron j in layer $l+1$ with relevance $R_j^{(l+1)}$, the relevance of neuron i in layer l is computed as:

$$R_i^{(l)} = \sum_j \frac{a_i^{(l)} w_{ij}^{(l,l+1)}}{\sum_{i'} a_{i'}^{(l)} w_{i'j}^{(l,l+1)}} R_j^{(l+1)} \tag{13}$$

where $a_i^{(l)}$ is the activation of neuron i in layer l, and $w_{ij}^{(l,l+1)}$ is the weight connecting neurons i and j.

Different LRP rules have been developed for different layer types:

LRP-ϵ rule: For avoiding numerical instabilities:

$$R_i^{(l)} = \sum_j \frac{a_i^{(l)} w_{ij}^{(l,l+1)}}{\sum_{i'} a_{i'}^{(l)} w_{i'j}^{(l,l+1)} + \epsilon \cdot sign(\sum_{i'} a_{i'}^{(l)} w_{i'j}^{(l,l+1)})} R_j^{(l+1)} \tag{14}$$

LRP-γ rule: For input layers to focus on positive evidence:

$$R_i^{(l)} = \sum_j \frac{a_i^{(l)} (w_{ij}^{(l,l+1)+} + \gamma w_{ij}^{(l,l+1)-})}{\sum_{i'} a_{i'}^{(l)} (w_{i'j}^{(l,l+1)+} + \gamma w_{i'j}^{(l,l+1)-})} R_j^{(l+1)} \tag{15}$$

where $w^+ = \max(0, w)$ and $w^- = \min(0, w)$.

For transformer-based models in mental health NLP, LRP can be applied to analyze attention mechanisms. The relevance propagation through multi-head attention can be computed as:

$$R_{i,k}^{(l)} = \sum_{h=1}^{H} \sum_{j=1}^{N} \frac{\alpha_{i,j}^{(h)} W_O^{(h)} W_V^{(h)} x_j}{\sum_{i'=1}^{N} \alpha_{i',j}^{(h)} W_O^{(h)} W_V^{(h)} x_j} R_{j,k}^{(l+1)} \tag{16}$$

where $\alpha_{i,j}^{(h)}$ is the attention weight between tokens i and j in head h, $W_O^{(h)}$ and $W_V^{(h)}$ are the output and value weight matrices, and H is the number of attention heads.

3.3 Advanced Gradient-Based Methods

3.3.1 SmoothGrad

SmoothGrad (Smilkov et al., 2017) reduces noise in gradient-based explanations by averaging gradients over multiple noisy versions of the input:

$$\hat{M}_c(x) = \frac{1}{n}\sum_{i=1}^{n} M_c(x + \mathcal{N}(0, \sigma^2)) \tag{17}$$

where $M_c(x)$ is the gradient-based attribution method, $\mathcal{N}(0, \sigma^2)$ is Gaussian noise, and n is the number of noisy samples.

3.3.2 GradCAM for Text

Adapted from computer vision, GradCAM for text computes the importance of different regions in the input text:

$$L^c_{GradCAM} = ReLU\left(\sum_k \alpha^c_k A^k\right) \tag{18}$$

where $\alpha^c_k = \frac{1}{Z}\sum_i \frac{\partial y^c}{\partial A^k_i}$ are the importance weights, A^k is the activation map of the k-th feature map, and y^c is the score for class c.

3.4 Model-Specific Methods

3.4.1 Attention Visualization and Analysis

For transformer-based models, attention weights provide insights into which tokens the model focuses on. The attention mechanism computes:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{19}$$

where Q, K, and V are the query, key, and value matrices, respectively.

However, attention weights may not always reflect true importance. Recent work has shown that attention weights can be manipulated without changing predictions, questioning their reliability as explanations.

Attention Rollout: Combines attention weights across layers to track information flow:

$$\tilde{A}^{(l)} = 0.5 \cdot A^{(l)} + 0.5 \cdot \tilde{A}^{(l-1)} A^{(l)} \tag{20}$$

where $A^{(l)}$ is the attention matrix at layer l.

Attention Flow: Tracks the flow of information through the network:

$$F^{(l)} = \frac{1}{2}(A^{(l)} + A^{(l)T}) \cdot F^{(l-1)} \tag{21}$$

In mental health applications, attention analysis has revealed that models learn to focus on emotional keywords.

3.4.2 Probing Studies

Probing studies train auxiliary classifiers on intermediate representations to understand what information is encoded. For a representation $h^{(l)}$ at layer l, a probing classifier g is trained to predict some linguistic property y:

$$\hat{y} = g(h^{(l)}) = W_{probe}h^{(l)} + b_{probe} \tag{22}$$

The accuracy of the probing classifier indicates how well the information is encoded at that layer.

Structural Probes: Specifically designed to understand syntactic structure:

$$d_{syntax}(w_i, w_j) = ||W_s h_i - W_s h_j||_2^2 \tag{23}$$

where W_s is a learned transformation matrix and d_{syntax} represents syntactic distance.

In mental health NLP, probing studies have revealed that models learn hierarchical representations of emotional and psychological states, with lower layers capturing basic emotional expressions and higher layers capturing complex psychological patterns.

3.5 Evaluation Metrics for Explanations

3.5.1 Faithfulness Metrics

Sufficiency: Measures how well the explanation alone can predict the model's output:

$$Sufficiency = \frac{1}{N}\sum_{i=1}^{N} \nVdash[f(x_i^{suff}) = f(x_i)] \tag{24}$$

where x_i^{suff} contains only the features deemed important by the explanation.

Comprehensiveness: Measures how much the prediction changes when important features are removed:

$$Comprehensiveness = \frac{1}{N}\sum_{i=1}^{N}|f(x_i)-f(x_i^{comp})| \tag{25}$$

where x_i^{comp} has important features removed.

3.5.2 Stability Metrics

Lipschitz Continuity: Measures explanation stability:

$$Lipschitz = \max_{x,x'} \frac{||E(x) - E(x')||}{||x - x'||} \tag{26}$$

where $E(x)$ is the explanation for input x.

These mathematical foundations provide the theoretical basis for understanding how different explainability methods work and their relative strengths and limitations in mental health NLP applications.

3.6 Recent Advances in Explainable Mental Health NLP

3.6.1 Counterfactual Explanations

Counterfactual explanations answer the question "What would need to change for the model to make a different prediction?" For text data, this involves finding minimal edits that flip the prediction:

$$x_{cf} = \arg\min_{x'} ||x - x'|| \text{ s.t. } f(x') \neq f(x) \tag{27}$$

In mental health applications, counterfactual explanations can help identify specific linguistic changes that would alter a depression or suicide risk assessment, providing actionable insights for intervention.

3.6.2 Concept-based Explanations

Testing with Concept Activation Vectors (TCAV) (Kim et al., 2018) measures the sensitivity of a model to human-interpretable concepts:

$$TCAV_{l,k,C} = \frac{1}{|X_k|}\sum_{x\in X_k} \nabla h_l(x) \cdot v_C \tag{28}$$

where v_C is the concept activation vector for concept C, $h_l(x)$ is the activation at layer l, and X_k is the set of examples of class k.

3.6.3 Adversarial Explanations

Adversarial examples can reveal model vulnerabilities in mental health applications. The perturbation is computed as:

$$x_{adv} = x + \epsilon \cdot sign(\nabla_x L(f(x), y)) \tag{29}$$

Understanding these vulnerabilities is crucial for ensuring robustness in clinical settings.

4 Applications and Case Studies

4.1 Depression Detection and Analysis

(Tadesse et al., 2019) developed an interpretable framework for depression detection using social media data. Their approach combined LIME explanations with clinical expertise to validate that the model was focusing on clinically relevant linguistic patterns rather than spurious correlations.

The study revealed that effective depression detection models focus on:

- Increased use of first-person singular pronouns
- Negative emotion words and expressions
- References to sleep disturbances and fatigue
- Social isolation indicators

However, LIME analysis also revealed concerning biases, such as the model associating certain demographic terms with depression risk, highlighting the importance of explainability for bias detection.

4.2 Empathetic Response Generation

(Sharma et al., 2022) developed a motivational virtual assistant that could generate empathetic responses for mental health support. They used attention visualization and gradient-based methods to understand how the model learned to generate appropriate empathetic responses.

The explainability analysis revealed:

- The model learned to attend to emotional cues in user messages

- Different attention heads specialized in different aspects of empathy (emotional, cognitive, compassionate)

- The generation process involved complex interactions between emotion detection and response strategy selection

4.3 Suicide Risk Assessment

Recent work on suicide risk assessment has emphasized the critical importance of explainability given the life-or-death nature of the application. (Zirikly et al., 2019) used SHAP values to analyze which linguistic features were most predictive of suicide risk, finding that:

- Direct expressions of suicidal ideation were not always the strongest predictors

- Indirect indicators such as hopelessness and social disconnection were often more informative

- Temporal patterns in posting behavior were as important as content

4.4 Multi-Modal Mental Health Analysis

Recent advances have explored combining textual data with other modalities for more comprehensive mental health assessment. (Park et al., 2024) demonstrated multimodal empathy detection using:

$$\begin{aligned} f_{\text{multimodal}} = {} & \alpha f_{\text{text}}(x_{\text{text}}) + \beta f_{\text{audio}}(x_{\text{audio}}) \\ & + \gamma f_{\text{visual}}(x_{\text{visual}}) \end{aligned} \tag{30}$$

where $\alpha + \beta + \gamma = 1$ and each f_i represents a modality-specific model.

Explainability in multimodal settings requires attribution across modalities:

$$\text{Attribution}_{\text{total}} = \sum_{m \in \{\text{text,audio,visual}\}} w_m \cdot \text{Attribution}_m \tag{31}$$

4.5 Longitudinal Mental Health Monitoring

Time-series analysis of mental health data requires explanations that capture temporal patterns. For RNN-based models, the hidden state evolution can be analyzed:

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t + b_h) \tag{32}$$

Temporal attribution methods compute the contribution of each time step:

$$A_t = \frac{\partial f(x_{1:T})}{\partial h_t} \cdot \frac{\partial h_t}{\partial x_t} \tag{33}$$

4.6 Personalized Mental Health Interventions

Personalized models adapt to individual linguistic patterns. Meta-learning approaches optimize for quick adaptation:

$$\theta_i = \theta - \alpha \nabla_\theta L_{task_i}(\theta) \tag{34}$$

where θ_i are personalized parameters for individual i.

Explanations must account for both global patterns and individual adaptations:

$$\begin{aligned} \text{Explanation}_i = {} & \text{GlobalExplanation} \\ & + \text{PersonalizedExplanation}_i \end{aligned} \tag{35}$$

5 Challenges and Limitations

5.1 Evaluation of Explainability

One of the fundamental challenges in explainable NLP for mental health is the lack of standardized evaluation metrics for explanation quality. Different stakeholders (clinicians, patients, researchers) may have different requirements for explanations.

Faithfulness vs. Plausibility: There is often a tension between explanations that accurately reflect model behavior (faithfulness) and explanations that make intuitive sense to humans (plausibility) (Jacovi and Goldberg, 2020).

Stability: Explanations should be stable across similar inputs, but many current methods produce inconsistent explanations for semantically similar texts.

5.2 Domain-Specific Challenges

Linguistic Variability: Mental health expressions vary significantly across demographics, cultures, and platforms, making it challenging to develop universally interpretable models.

Temporal Dynamics: Mental health states change over time, but most current explainability methods focus on static snapshots rather than temporal patterns.

Multimodal Data: Increasingly, mental health applications incorporate multiple modalities (text, audio, visual), but explainability methods for multimodal models are still limited.

5.3 Ethical Considerations

Privacy: Explainability techniques may inadvertently reveal sensitive information about individuals or groups in the training data.

Stigma: Explanations that reinforce mental health stigma or stereotypes can be harmful even if they accurately reflect model behavior.

Over-reliance: There is a risk that clinicians may over-rely on AI explanations rather than using them as supplementary information.

6 Future Directions and Opportunities

6.1 Human-Centered Explainability

Future research should focus on developing explanation techniques that are tailored to the specific needs and expertise of different stakeholders in mental health care:

Clinician-Focused Explanations: Explanations should align with clinical reasoning processes and use familiar terminology and concepts.

Patient-Facing Explanations: Patients should be able to understand how AI systems analyze their data and make recommendations about their care.

Researcher Explanations: Researchers need detailed technical explanations to validate models and identify areas for improvement.

6.2 Causal Explainability

Moving beyond correlation-based explanations to causal explanations would provide deeper insights into mental health phenomena. This involves:

- Developing causal models of mental health conditions
- Integrating domain knowledge into explainability methods
- Using counterfactual reasoning to understand intervention effects

6.3 Multimodal and Temporal Explainability

As mental health NLP systems become more sophisticated, explainability methods must evolve to handle:

- Multimodal inputs (text, speech, physiological signals)
- Temporal patterns and longitudinal data
- Context-dependent explanations

6.4 Personalized Explainability

Different individuals may require different types of explanations based on their background, preferences, and mental health conditions. Developing personalized explanation systems could improve user understanding and trust.

6.5 Regulatory and Standardization Efforts

The field would benefit from:

- Standardized evaluation metrics for explanation quality
- Guidelines for explainable AI in mental health applications
- Regulatory frameworks that balance transparency with privacy

7 Conclusion

This survey has provided a comprehensive overview of explainability techniques in NLP for mental health applications. We have examined various explainability methods including LIME, SHAP, influence functions, gradient-based approaches, and LRP, analyzing their applications in depression detection, suicide risk assessment, empathetic dialogue systems, and clinical decision support.

Key findings from our review include:

1. **Method Diversity:** Different explainability methods provide complementary insights, and multi-method approaches often yield the most comprehensive understanding.

2. **Domain Specificity:** Mental health applications have unique requirements for explainability due to the high-stakes nature of decisions and the complexity of mental health phenomena.

3. **Stakeholder Needs:** Different stakeholders (clinicians, patients, researchers) require different types of explanations, highlighting the need for human-centered design.

4. **Bias Detection:** Explainability methods are crucial for identifying and mitigating biases in mental health NLP systems.

5. **Trust and Adoption:** Interpretable systems show higher adoption rates among healthcare professionals, emphasizing the practical importance of explainability.

Despite significant progress, several challenges remain, including the lack of standardized evaluation metrics, the complexity of temporal and multimodal explanations, and ethical considerations around privacy and stigma. Future research should focus on developing human-centered explainability approaches, incorporating causal reasoning, and establishing regulatory frameworks for explainable AI in mental health.

As NLP systems become increasingly integrated into mental health care, the development of trustworthy, interpretable, and effective explainability methods will be crucial for realizing the full potential of AI in supporting mental health and well-being.

References

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2019. Ethical research in machine learning for health care. *Nature Machine Intelligence*, 1(9):377–383.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *npj Digital Medicine*, 3(1):1–11.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 51–60.

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights*, 10:1178222618792860.

Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. 2017. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205. Association for Computational Linguistics.

Shaoxiong Ji, Celina Ping Yu, Sai-fu Fung, Shirui Pan, and Pekka Marttinen. 2021. Suicidal ideation detection: A review of machine learning methods and applications. *IEEE Transactions on Computational Social Systems*, 8(1):214–226.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viégas, and Rory Sayres. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 2668–2677. PMLR.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. *International conference on machine learning*, pages 1885–1893.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Wei Luo, Dinh Phung, Truyen Tran, Sunil Gupta, Santu Rana, Chandan Karmakar, Alistair Shilton, John Yearwood, Nevenka Dimitrova, Tu Bao Ho, et al. 2016. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *Journal of medical Internet research*, 18(12):e323.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Mime: Mimicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 8968–8979.

Youngjae Park, Hyungtae Lim, Heuiseok Kim, and Jooyoung Yeo. 2024. Empathy through multimodality in conversational interfaces. *arXiv preprint arXiv:2405.04777*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2022. Computational approaches to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2392–2407.

Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, and Wenwu Zhu. 2017. Depression detection via harvesting social media: A multimodal dictionary learning solution. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 3838–3844.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. *International conference on machine learning*, pages 3319–3328.

Michael Mesfin Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2019. Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7:44883–44893.

Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. 2019. What clinicians want: contextualizing explainable machine learning for clinical end use. *Proceedings of the machine learning for healthcare conference*, pages 359–380.

David S Watson, Jenny Krutzinna, Ian N Bruce, Christopher EM Griffiths, Iain B McInnes, Michael R Barnes, and Luciano Floridi. 2019. Clinical applications of machine learning algorithms: beyond the black box. *BMJ*, 364.

Ayah Zirikly, Philip Resnik, Ozlem Uzuner, and Kristy Hollingshead. 2019. Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33.