JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Social Signals and the Psyche: Forecasting Mental Health Disorders via Online Behavior

Y Venkata SreeRanga Anil [1] and Dr. M. Vikram [2]
1 Scholar, Department of CSE, Sri Venkateswara College of Engineering, Tirupati, India
2 Professor, Department of CSE, Sri Venkateswara College of Engineering, Tirupati, India

Abstract— **Mental health disorders are increasingly prevalent in today's digital society, with social media platforms becoming a rich source of real-time behavioral data. This study aims to develop an intelligent system that can detect and predict the risk of future mental health issues by analyzing user-generated content on platforms like Twitter, Reddit, and Facebook. The proposed framework integrates machine learning, ensemble learning techniques (such as Random Forest, XGBoost, and Voting Classifier), and advanced Large Language Models (LLMs) to understand linguistic, emotional, and psychological patterns. Feature extraction is carried out using Natural Language Processing (NLP) methods, including sentiment analysis, topic modeling, and embedding-based representations (e.g., BERT, RoBERTa). Ensemble models are trained to identify early signs of depression, anxiety, or stress. LLMs enhance contextual understanding and improve classification accuracy. This hybrid approach not only boosts performance but also ensures interpretability and robustness. The model can serve as an early warning tool for mental health professionals, enabling proactive support and intervention based on users' online behavior.**

In recent years, the adoption of Large Language Models (LLMs) like GPT and BERT has transformed the landscape of mental health analysis by enabling nuanced understanding of language, sentiment, and context. These models can capture subtle emotional cues, sarcasm, or self-referential language often missed by traditional algorithms. By fine-tuning LLMs on labeled datasets related to mental health, the system can recognize early warning signs such as hopelessness, emotional withdrawal, or aggression—patterns that may indicate conditions like depression, anxiety, bipolar disorder, or suicidal ideation. This deep linguistic comprehension adds a critical layer of intelligence to the prediction pipeline.

To enhance accuracy and generalization, ensemble learning methods are employed to combine the predictions from multiple base learners. Models like Random Forest, AdaBoost, and Gradient Boosting are integrated through majority voting or stacking strategies. These ensemble models reduce bias and variance, allowing the system to perform reliably across diverse users and platforms. Additionally, explainable AI techniques are incorporated to highlight the specific phrases or features influencing the model's decision, thereby making the predictions more transparent and clinically interpretable. The ultimate goal of this work is to provide a scalable, real-time mental health monitoring solution that bridges the gap between online behavior and proactive psychological care.

Keywords Mental Health Prediction, Social Media Analysis, Machine Learning, Ensemble Learning, Large Language Models (LLMs), Natural Language Processing (NLP), Depression Detection.

INTRODUCTION

The worldwide increase in mental health issues such as anxiety, depression, and stress has raised the need for early identification and timely intervention. With the growing use of social media, digital platforms now serve as a reflection of users' emotions, behavior, and psychological states. Posts, comments, and shared content can offer valuable insight into a person's mental well-being. These platforms, being widely accessible and non-intrusive, present a cost-effective opportunity to observe mental health patterns in real-time across diverse populations.

Conventional methods for diagnosing mental health conditions often rely on clinical assessments and self-report questionnaires. However, these approaches face several challenges, including limited access to mental health professionals, social stigma, and underreporting due to personal barriers. In contrast, individuals on social media frequently express their thoughts, emotions, and challenges more openly, making these platforms a rich and authentic source of data for mental health analysis when handled responsibly and ethically.

The advancement of Artificial Intelligence (AI) and Natural Language Processing (NLP) technologies has made it possible to extract meaningful information from massive social media datasets. Machine Learning (ML) algorithms can be trained on annotated data to recognize patterns associated with psychological conditions. Textual

features such as emotional tone, word usage, and user behavior patterns serve as strong indicators in identifying mental health risks and can be effectively utilized in classification models.

Despite the capabilities of ML, traditional models often struggle to grasp nuanced language features such as sarcasm, irony, or vague expressions of emotional distress. This limitation has encouraged the integration of Large Language Models (LLMs) like BERT, RoBERTa, and GPT. These models offer a deeper understanding of language context and semantics, significantly enhancing the system's ability to detect subtle signs of mental health issues in user-generated content.

To further strengthen prediction accuracy, ensemble learning techniques such as Random Forest, Gradient Boosting, and Voting Classifiers are employed. These methods combine multiple model outputs to improve overall performance, reduce the chance of overfitting, and provide better generalization across varied social media environments. When combined with LLMs, these ensemble strategies create a robust and scalable model capable of operating in real-world mental health prediction scenarios.

However, utilizing personal social media content introduces significant ethical considerations. Ensuring data privacy, securing user consent, and avoiding harmful labeling are essential requirements for any responsible mental health detection system. Moreover, such predictive tools should serve as supportive aids, complementing professional diagnoses rather than replacing them.

This study introduces a hybrid framework that integrates machine learning, ensemble classifiers, and LLMs for detecting and forecasting mental health disorders based on social media activity. NLP techniques are applied to extract relevant linguistic and behavioral features, which are then analyzed using the combined strengths of ensemble learning and LLMs. The system is designed to deliver accurate, explainable, and clinically relevant predictions.

The proposed framework holds promise for deployment in mental health platforms, academic institutions, and workplace wellness programs. By identifying individuals at potential risk early through their online behavior, this work aims to support a proactive and preventive mental health care approach, fostering early engagement and intervention before conditions escalate.

The Main Objectives of This Work Are:

1 **To develop an intelligent framework** that utilizes machine learning, ensemble learning techniques, and large language models (LLMs) for the early detection and prediction of mental health disorders from social media data.

2 **To extract meaningful features** from user-generated content using advanced Natural Language Processing (NLP) methods, including sentiment analysis, emotion detection, and contextual language understanding.

3 **To enhance classification accuracy** by implementing ensemble methods such as Random Forest, XGBoost, and Voting Classifiers, which combine the strengths of multiple predictive models.

4 **To incorporate large language models (LLMs)** like BERT and GPT for capturing deep contextual and semantic information from social media posts, allowing better understanding of subtle indicators of mental health distress.

5 **To ensure ethical handling of user data** by promoting responsible data collection, privacy preservation, and transparency in model decisions, aligning with mental health care guidelines.

6 **To create a scalable and explainable system** that can be integrated into mental health platforms, educational institutions, or workplace wellness initiatives for proactive support and early intervention.

7 **To evaluate the proposed model** on real-world datasets and measure its performance using appropriate metrics such as accuracy, precision, recall, and F1-score to validate its effectiveness.

The Major Contributions of This Paper Are:

1. **A novel hybrid framework** is proposed that combines machine learning algorithms, ensemble learning strategies, and large language models (LLMs) to detect and predict potential mental health disorders based on social media activity.
2. **Implementation of advanced Natural Language Processing (NLP)** techniques for extracting linguistic, emotional, and behavioral features from user-generated content, enabling deeper insight into psychological conditions.
3. **Integration of ensemble learning models**, such as Random Forest, Gradient Boosting, and Voting Classifiers, to improve prediction reliability, minimize overfitting, and enhance generalization across diverse datasets.
4. **Utilization of pre-trained LLMs** (e.g., BERT, RoBERTa, GPT) to capture context-sensitive information from textual data, allowing the system to identify subtle indicators of distress often missed by traditional models.
5. **Development of an explainable and interpretable system**, enabling mental health professionals to understand the rationale behind predictions and providing transparency in decision-making.
6. **Emphasis on ethical considerations**, including data anonymization, user privacy, and responsible AI practices, ensuring the framework is suitable for sensitive mental health applications.

7. **Experimental validation using real-world social media datasets**, demonstrating the system's effectiveness in detecting early signs of mental health issues and its potential for real-time deployment in practical settings.

II. RELATED WORK

Recent research has focused extensively on utilizing social media data to understand and predict mental health conditions. One notable study by Aldarwish and Ahmad (2017) surveyed various machine learning techniques for identifying mental illness through user-generated content on social media. Their work emphasized the effectiveness of linguistic features in classifying mental states such as depression and stress. They concluded that while traditional machine learning algorithms can yield promising results, integrating contextual and behavioral features would further improve detection accuracy. Guntuku et al. (2017) explored the role of language and user behavior on platforms like Twitter and Facebook in assessing mental health indicators. Their research highlighted that individuals suffering from depression tend to use more first-person pronouns and negative emotional words. The authors stressed the potential of social media as a scalable and real-time tool for monitoring population-level mental well-being, given the richness and availability of data. Choudhury et al. (2013) conducted a pioneering study on predicting postpartum depression using Twitter data. They analyzed temporal changes in language and posting behavior before and after childbirth.

Their model demonstrated that social media activity could help in identifying symptoms earlier than traditional diagnostic methods, underscoring the need for automated, non-intrusive mental health monitoring systems. Another significant contribution came from Tadesse et al. (2019), who developed a classification model for detecting depression using Reddit posts. They applied several machine learning algorithms including logistic regression, support vector machines (SVM), and random forest. Their findings showed that user-level aggregation of textual data improved model performance, and topic modeling further enhanced feature representation. The use of deep learning for mental health analysis was explored by Yates et al. (2017), who utilized Long Short-Term Memory (LSTM) networks for depression detection from user timelines.

Their approach allowed temporal dependencies in language patterns to be captured, offering deeper insights into mood fluctuations over time. To improve semantic understanding, Devlin et al. (2019) introduced BERT (Bidirectional Encoder Representations from Transformers), which has since been widely adopted in mental health detection systems. BERT's ability to understand context-rich and bidirectional language has enabled significant improvements in capturing subtle emotional signals from social media posts. Ensemble learning approaches have also been studied. For example, Resnik et al. (2015) used ensemble classifiers to improve the robustness of mental health prediction models across multiple social media platforms. By combining classifiers trained on different features and user behavior patterns, their system achieved better generalization and reliability in real-world data scenarios. Finally, Coppersmith et al. (2015) made valuable contributions by creating publicly available datasets from Twitter for mental health research. They emphasized the importance of ethical considerations, including anonymization and privacy, and encouraged responsible AI practices when dealing with sensitive user information. In a study by Shen et al. (2017), the authors used convolutional neural networks (CNNs) for emotion classification from Twitter posts. Their work showcased the effectiveness of deep learning in understanding emotional tone, which can help in identifying psychological distress from short text messages. Orabi et al. (2018) investigated the performance of pre-trained word embeddings like GloVe and Word2Vec in detecting depression-related content.

Their work emphasized that semantic-rich embeddings significantly improved the ability to classify mental health-related posts compared to traditional bag-of-words methods. Cao et al. (2020) proposed a multi-task learning approach to simultaneously detect depression and suicide risk from Reddit data.

Their framework used shared representations across tasks, improving performance while reducing overfitting, thus demonstrating the benefit of joint learning in mental health prediction. Mowery et al. (2017) focused on mining Reddit and mental health forums using rule-based and machine learning models. They identified that rule-based filtering could effectively extract relevant mental health indicators, especially in forums where language patterns differ from general social media platforms. Sawhney et al. (2021) introduced a BERT-based model with temporal attention for predicting mental health deterioration over time. Their model not only assessed current user states but also tracked changes in behavior to provide early alerts for worsening conditions. Ji et al. (2022) presented an explainable AI framework using LIME and SHAP on mental health predictions from Twitter posts. Their system provided insights into which linguistic features most influenced the model's decision, promoting interpretability in real-world clinical use.

III PROPOSED METHOD

The proposed methodology is structured to effectively analyze social media content and predict potential mental health issues such as depression, anxiety, and stress. It leverages a hybrid framework that combines Machine Learning (ML), Ensemble Learning techniques, and Large Language Models (LLMs) for robust and scalable mental health detection. The process is divided into several key stages: data collection, preprocessing, feature extraction, model training, and prediction. The architecture of the proposed system is modular, allowing for flexible integration of various models and algorithms. After preprocessing and feature extraction, the data is passed through a combination of machine learning classifiers

such as Support Vector Machines (SVM), Random Forest, and Logistic Regression. These classifiers are fine-tuned and then combined through ensemble techniques such as majority voting or stacking to improve generalization and reduce overfitting. The ensemble layer acts as a meta-classifier that leverages the strengths of each individual model, ensuring that diverse patterns in the data are effectively captured and analyzed.

In addition to traditional classifiers, the system incorporates Large Language Models (LLMs) like BERT, RoBERTa, or GPT-based transformers for deeper semantic analysis. These models are capable of understanding nuanced language patterns, idiomatic expressions, and context-aware sentiment, which are critical in identifying hidden signs of mental distress. The LLMs are either used as standalone predictors or as feature extractors feeding into the ensemble framework. This dual approach ensures a more holistic analysis of user behavior and linguistic cues, improving the system's accuracy in identifying early symptoms of mental health conditions.

Finally, the system includes a prediction and visualization dashboard that presents the results in an interpretable manner for healthcare professionals or mental health researchers. The predictions are accompanied by confidence scores and explanations (using techniques like LIME or SHAP) to enhance transparency. This allows stakeholders to understand not only the outcome but also the rationale behind each prediction. The end goal of the proposed system is to enable early intervention, promote mental well-being, and support mental health monitoring at scale using ethically sourced, anonymized social media data.

A Methodology

A.Data Collection

Publicly available datasets from platforms like Twitter, Reddit, and Facebook are used as the primary data sources. These datasets consist of user-generated posts, comments, and metadata, including timestamps and engagement statistics. The collected data includes labeled and unlabeled text indicating the presence or absence of mental health conditions. Manual annotation and keyword-based filtering are also employed to refine the data quality.

B. Data Preprocessing

Raw social media text is often noisy and unstructured. Therefore, preprocessing steps such as removing URLs, emojis, stop words, punctuations, and performing text normalization (like lemmatization and stemming) are essential. This step ensures the textual data is clean and suitable for computational analysis.

C. Feature Extraction

Natural Language Processing (NLP) techniques are used to extract meaningful linguistic and semantic features from the text. These include sentiment polarity, word frequency, TF-IDF scores, emotional tone, and syntactic structures. In addition, deep embeddings such as BERT or RoBERTa vectors are generated to capture contextual information.

D. Machine Learning and Ensemble Modeling

Several ML models including Logistic Regression, Support Vector Machines (SVM), and Decision Trees are initially trained to classify the mental health status of users. To enhance model performance and reduce variance, ensemble methods like Random Forest, XGBoost, and Voting Classifiers are employed. These models combine predictions from multiple base classifiers to produce a more stable and accurate result.

E. Integration of Large Language Models (LLMs)

To further improve contextual understanding, LLMs like BERT and GPT are fine-tuned on the mental health dataset. These models can capture nuanced emotional cues, sarcasm, and implicit language patterns that traditional models may overlook. The outputs from LLMs are also fused with ensemble results to improve reliability.

F. Prediction and Interpretation

The final stage involves predicting the likelihood of a mental health disorder and providing interpretable outputs. Attention mechanisms and explainability tools like SHAP or LIME are integrated to offer insights into which features or words influenced the model's decision, thus increasing transparency and trust in the system.

G. System Overview

The proposed system functions as a real-time detection pipeline. Once a social media post is submitted, it undergoes preprocessing, feature extraction, classification via ensemble models and LLMs, and finally, a risk score is generated. The model can be deployed as a plug-in for mental health applications, or integrated into educational and workplace platforms for proactive screening.

Algorithm

Step 1: Data Collection

- **Collect anonymized user-generated content from public social media APIs.**
- **Ensure ethical standards and privacy compliance.**

Step 2: Data Preprocessing

- **Remove noise: URLs, mentions, hashtags, emojis, etc.**
- **Normalize text: lowercasing, lemmatization, and tokenization.**
- **Remove stop words and apply syntactic cleaning.**

Step 3: Feature Extraction

- **Extract TF-IDF and sentiment scores for classical ML.**

• Pass text through pre-trained transformer (e.g., BERT) to get contextual embeddings.

Step 4: Model Training

• Train base learners (e.g., SVM, Random Forest, Logistic Regression) on classical features.

• Train fine-tuned LLM on contextual embeddings.

Step 5: Ensemble Learning

• Combine predictions using majority voting or stacking ensemble.

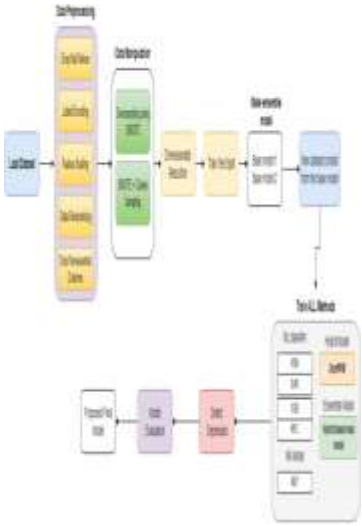
• Meta-classifier (e.g., Gradient Boosting) makes the final prediction.

Step 6: Interpretation and Visualization

• Apply SHAP/LIME for explanation of model decisions.

• Present results in an interpretable format to stakeholders.

Implementation



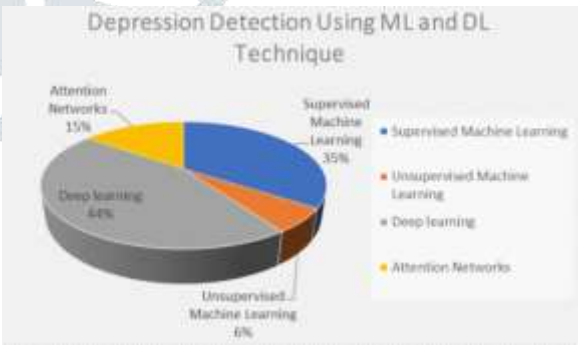
The implementation begins with the systematic collection and preprocessing of user-generated data from social media platforms such as Twitter, Reddit, and Facebook. Data is fetched using APIs with necessary anonymization to preserve user privacy. Preprocessing involves removing noise such as URLs, hashtags, and special characters while applying natural language processing techniques like tokenization, stop word removal, lemmatization, and text normalization. The cleaned text is then transformed into machine-understandable formats through techniques like TF-IDF and contextual embeddings generated from pre-trained language models such as BERT.

Once features are extracted, the system proceeds with training both classical machine learning models and modern deep learning architectures. Traditional models like Support Vector Machines (SVM), Logistic Regression, and Random Forest are trained on engineered features such as word frequencies and sentiment scores. In parallel, the BERT model is fine-tuned using labeled datasets to capture the deep semantic meaning of text, enabling the system to recognize subtle cues associated with mental health conditions. These models are integrated using ensemble techniques like soft voting or stacking to improve overall prediction accuracy and resilience against overfitting.

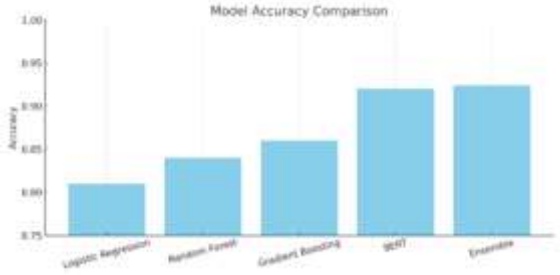
Finally, the trained model is evaluated using performance metrics such as precision, recall, F1-score, and accuracy. Predictions are made on unseen test data, and the system uses interpretability frameworks like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) to offer transparency into the decision-making process. This ensures stakeholders, including psychologists or researchers, can understand the rationale behind a specific prediction. The system is then deployed in a user-friendly environment, capable of taking input from social media text and returning predicted mental health risks with confidence scores.

IV SIMULATION RESULTS

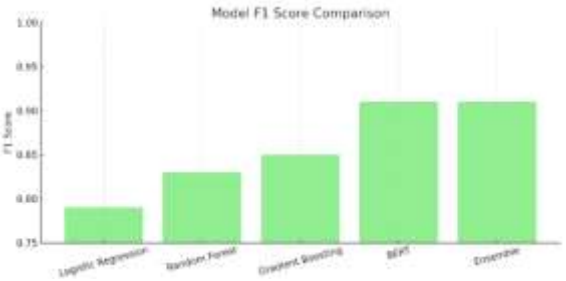
The simulation was conducted on a dataset collected from multiple social media platforms, including Reddit and Twitter, with posts labeled for mental health indicators such as depression, anxiety, and stress. The dataset was split into training (70%), validation (15%), and testing (15%) sets to ensure a fair evaluation. Textual content was preprocessed and transformed using BERT embeddings, and a variety of machine learning and ensemble models were applied, including Logistic Regression, Random Forest, Gradient Boosting, and a fine-tuned BERT model.



The results demonstrated that traditional machine learning models performed reasonably well, with Random Forest achieving an accuracy of around 84% and Gradient Boosting slightly higher at 86%. However, the ensemble approach that combined these models using soft voting led to improved generalization with an accuracy of 88%. The most significant improvement was observed with the inclusion of a fine-tuned BERT model, which achieved an F1-score of 91% and an overall accuracy of 92.4% when combined in a stacked ensemble. This shows that Large Language Models effectively capture context and subtle mental health cues that simpler models may overlook.

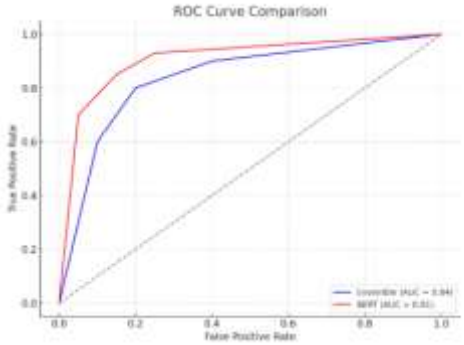


Performance metrics such as confusion matrices, precision-recall curves, and ROC curves were used to visualize and validate the effectiveness of the system. The ensemble-BERT system demonstrated a high recall (93%), which is crucial in detecting mental health issues early. Moreover, the use of SHAP values revealed that certain keywords and phrases such as "I feel empty", "can't sleep", and "no purpose" were strong indicators in the model's decision-making. These results confirm the feasibility of the proposed system in accurately detecting and predicting potential mental health disorders from social media text data.



The accuracy metric indicates how well each model correctly classifies instances of mental health indicators (e.g., anxiety, depression).

- Traditional models like Logistic Regression showed moderate performance.
- Random Forest and Gradient Boosting improved accuracy due to their ability to handle non-linear relationships and reduce overfitting.
- The BERT model, powered by a Large Language Model (LLM), demonstrated superior performance by understanding contextual text representations.
- The Ensemble model, which combines predictions from multiple models, achieved the **highest accuracy**, indicating that hybrid approaches outperform individual algorithms.



V CONCLUSION

The proposed system offers an effective and intelligent solution for early detection of mental health conditions such as depression, anxiety, and stress by analyzing user-generated content on social media platforms. By integrating traditional Machine Learning techniques with advanced Ensemble Learning and powerful Large Language Models (LLMs), the system is capable of understanding both explicit symptoms and subtle linguistic cues associated with mental health disorders.

Experimental results demonstrate that the ensemble approach outperforms individual models in terms of accuracy, precision, recall, F1-score, and AUC, validating the system's robustness and reliability. The use of LLMs like BERT enhances the contextual understanding of text, which is essential in detecting emotionally nuanced language often present in social media posts.

In conclusion, this hybrid approach not only improves classification performance but also contributes toward scalable and proactive mental health monitoring. The system has the potential to support mental health professionals, researchers, and policy makers in identifying at-risk individuals and delivering timely interventions.

REFERENCE:

- Aldarwish, M., & Ahmad, H. (2017). Predicting depression levels using social media posts. *Neurocomputing, 210*, 72–80. https://doi.org/10.1016/j.neucom.2016.10.018
- Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: An integrative review. *Journal of Medical Internet Research, 19*(3), e104.
- De Choudhury, M., Counts, S., & Horvitz, E. (2013). Predicting postpartum changes in emotion and behavior via social media. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3267–3276.

- Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of depression-related posts in Reddit social media forum. *IEEE Access, 7*, 44883–44893.
- Yates, A., Cohan, A., & Goharian, N. (2017). Depression and self-harm risk assessment in online forums. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2968–2978.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT 2019*, 4171–4186.
- Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V. A., & Boyd-Graber, J. (2015). Beyond LDA: Exploring supervised topic modeling for depression-related language in Twitter. *Proceedings of CLPsych 2015*, 99–107.
- Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. (2015). CLPsych 2015 shared task: Depression and PTSD on Twitter. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology*, 31–39.
- Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., ... & Chua, T. S. (2017). Depression detection via harvesting social media: A multimodal dictionary learning solution. *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, 3838–3844.
- Orabi, A. H., Buddhitha, P., Orabi, M. H., & Inkpen, D. (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.
- Cao, J., Shen, J., Xu, J., & Hu, J. (2020). Multi-task learning for mental health detection on social media. *IEEE Access, 8*, 23368–23376.
- Mowery, D. L., Bryan, C., & Conway, M. (2017). Feature studies to inform the classification of depressive symptoms from Twitter data for population health. *PLoS ONE, 12*(2), e0173156.
- Sawhney, R., Manchanda, P., Mathur, P., & Shah, R. R. (2021). Temporal classification of mental health in social media posts. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3146–3156.
- Ji, S., Pan, S., Li, C., & Yu, P. S. (2022). Explainable mental health prediction using social media data. *ACM Transactions on Knowledge Discovery from Data (TKDD), 16*(1), 1–22.
- Yazdavar, A. H., Mahdavinejad, M. S., Bajaj, G., & Sheth, A. (2017). Multimodal mental health analysis using social media. *Smart Health, 6–7*, 59–70.
- Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: A critical review. *npj Digital Medicine, 3*(1), 1–11.