Reddit social media text analysis for depression prediction: using logistic regression with enhanced term frequency-inverse document frequency features

Madan Mohan Tito Ayyalasomayajula[1], Akshay Agarwal[2], Shahnawaz Khan[3]
[1]Department of Computer Science, School of Business and Technology, Aspen University, Phoenix, United States
[2]Department of Computer Science, Illinois Institute of Technology, Chicago, United States
[3]Department of computer Science, Bundelkhand University, Jhansi, India

Article Info

Article history:

Received May 10, 2024
Revised Jul 11, 2024
Accepted Jul 17, 2024

Keywords:

Language
Language analysis
Machine learning
Mental health
Reddit
Term frequency-depression
Term frequency-inverse document frequency

ABSTRACT

Language provides significant insights into an individual's emotional state, social status, and personality traits. This research aims to enhance depression detection through the analysis of linguistic features and various dataset attributes. The dataset, sourced from the social networking platform Reddit, comprises posts and comments from individuals diagnosed with depression. Logistic regression with term frequency-inverse document frequency (TF-IDF) is employed as the primary model for text classification. To improve model performance, a novel feature—the average time interval between consecutive posts or comments—is introduced, contributing to a marginal but noteworthy improvement in accuracy. The proposed model demonstrates superior F1 scores compared to other models applied to the same dataset. Given the increasing recognition of mental health's significance, accurately diagnosing mental disorders is of paramount importance. This study underscores the potential of leveraging linguistic analysis and advanced machine learning techniques to identify depressive symptoms, thereby contributing to more effective mental health diagnostics and interventions.

Corresponding Author:

Madan Mohan Tito Ayyalasomayajula
Department of Computer Science, School of Business and Technology, Aspen University
Phoenix, Arizona, United States
Email: mail2tito@gmail.com

1. INTRODUCTION

Depression constitutes a prevalent psychiatric condition, often underestimated in significance. Inadequate management may result in severe complications, functional impairment, psychotic manifestations, and in extreme instances suicidal behavior. World Health Organization (WHO) data indicate depression affects more than 264 million individuals globally [1]. As with numerous mental disorders, early identification potentially facilitates preventive interventions.

The significance of depression detection and the role of social media in addressing this challenge is paramount in contemporary healthcare. Early detection of depression is crucial for timely diagnosis and treatment, reducing associated risks. Current detection methods relying on self-report questionnaires are costly, time-consuming, and prone to cognitive biases. However, people's evolving lifestyles pose challenges for universal depression detection models, and this highlights the need for low-cost, effective passive methods. Leveraging social media platforms for automatic detection and understanding of mental states offers advantages over manual methods, avoiding recall bias and revealing unique correlations between digital behaviors and depression symptoms.

Technological advancements increase online engagement, which leads to more online presence, with digital behavior on social platforms and websites potentially yielding substantial insights into individual characteristics. Linguistic patterns serve as robust indicators of personality, emotional state, social standing, and mental well-being. Unsurprisingly, individuals experiencing depression frequently employ vocabulary conveying negative affect, particularly pessimistic adjectives and adverbs such as "lonely" and "sad." Notably, depressed persons' heightened usage of first-person singular pronouns such as “myself,” “me,” and “I” suggests increased self-focus [2]. This linguistic tendency provides valuable diagnostic information regarding mental state and cognitive patterns associated with depressive disorders.

The advent of artificial intelligence (AI) powered systems that support decisions has transformed strategy planning and implementation in several fields [3]. In the current digitally linked society, it is critical to recognize and treat mental health issues. Academics have started working on groundbreaking projects to understand depression via social media sites like Reddit using natural language processing and machine learning. This research explores the relationship between mental health analysis and AI-driven decision support systems, with a particular emphasis on the novel method of identifying depressed people by examining Reddit language patterns.

Social media content analysis often yields valuable insights into mental states. Current depression-related technologies operate reactively to treat depression disorders. Digital user monitoring or digital tracking identifies certain risks, triggering alerts upon illness manifestation or unethical conduct [4]. Natural language processing (NLP) application to typical social network posts offers superior early detection potential [5]. This investigation proposes an NLP-based methodology for assessing depressive ideation, thoughts, and intentions. The approach extracts phrases predominantly utilized by Reddit users exhibiting depressive tendencies alongside group-specific features. The dataset originally reported by Losada and Crestani [6] forms the foundation of this analysis.

2. RELATED WORKS AND PROBLEM STATEMENT

Zaghouani [7] analyzed extensive social media data for youth depression detection. The study proposed developing a sentiment analysis-based linguistically annotated corpus examining adolescent online behavior in the Middle East and North Africa (MENA). Goal: establish a broad user base with precise self-reported depression indicators. Vij and Pruthi [8] explored automated psychometric analyzers utilizing sentiment analysis and emotion recognition in healthcare. Objective: create a self-service medical kiosk with rapid computational linguistics capabilities, generating concise emotional health summaries based on patient history. Almouzini *et al.* [9] identified depressed Arabic Twitter users. Developed prediction model via Arabic sentiment analysis and supervised learning. Finding: depressed individuals exhibit increased social isolation. Priya *et al.* [10] applied ML algorithms for stress, depression, and anxiety prediction. The study evaluated five depression levels using ML algorithms. Random forest achieved the highest accuracy (91%, 89%). Feuston and Piper [11] explored how mental illness is expressed on Instagram through manual post collection, interviews, and visual methodologies. The essayists find users negotiate mental health visibility with blurred boundaries between health and illness, highlighting reposting and remixing as essential participatory forms. Murnane *et al.* [12] proposed developing technology for long-term cognitive health management and social ecosystems. Researchers focused on the patient’s point of view, proposing design concepts for collaborative informatics infrastructures and interfaces supporting personal data activities within social ecologies. Pater *et al.* [13] investigated a situation in which eating disorder patients employed digital self-harm indicators. Future research might compare post-intervention data to pre-intervention data to determine changes in patients’ online identity presentations. Xu *et al.* [14] proposed using contextually filtered attributes and routing behavior for detecting depression in college students. This research describes a new rule mining-based technique for automatically synthesizing contextually filtered features that outperform current feature selection techniques for depression diagnosis.

Trifan *et al.* [15] identified psycholinguistic characteristics in social media writings that help us comprehend depression. The authors were enthusiastic about exploring other psycholinguistic components with people who could shed light on them via clinical papers in a future study. Mathur *et al.* [16] proposed utilizing temporal psycholinguistic cues to assess suicide intention. This study aims to fill a vacuum in research by using both qualitative and quantitative research methodologies to examine the effects of enhancing the identification of suicidal thoughts represented in written language. Losada and Crestani's study [6] made valuable contributions to the subject by introducing the early risk detection error (ERDE) metrics as a new assessment statistic. This metric explicitly measures the speed and accuracy of spotting positive circumstances. Wang *et al.* [17] employed sentiment analysis to identify user depression. Word and artificial rules should decide each micro-depressive blog’s inclination. Using the provided approach and ten psychologically confirmed depressed traits, a depression detection framework is created.

Song *et al.* [18] demonstrated another excellent Japanese population mental health assessment method. They developed neural multi-task learning (MTL) models for nine prediction tasks. The research found that choosing the MTL and supplemental activities for a mental state can enhance performance.

2.1. Problem statement

In the review of existing literature on depression, several assertions regarding the characteristics of depressed individuals and their engagement on social media have surfaced. These assertions encompass using absolute terms, the expression of negative emotions in tweets, user posting frequency, and intervals between posts. A notable finding concerns the timing of posts among users exhibiting depressive symptoms. However, these claims were previously explored in 2019 by Banovic *et al.* [19], where they were deemed insignificant within the context of the dataset analyzed. The study aims further to investigate these features and behaviors in this domain. The data analysis has revealed that while absolutist terms (e.g., absolutely, entirely, completely) show no significant differences between depressed and non-depressed groups, average words per post and the number of comments per user exhibit meaningful distinctions [20]. This underscores the importance of examining various behavioral indicators to distinguish between these groups.

3. PROPOSED METHOD

3.1. Dataset description

The dataset encompasses Reddit posts and comments from 892 distinct users, providing a comprehensive sample for analysis. Among these, 137 individuals are undergoing depression treatment, while the remainder serve as a control group, offering a balanced perspective. Reddit API limitation becomes apparent: 1000 posts and 1000 comments per user, as depicted in Figure 1, potentially influencing data collection methodology. Posts and comments are chronologically arranged, a feature essential for early depression identification and temporal analysis of linguistic patterns. Depressive classification criteria adhere to strict guidelines: public acknowledgment of depression diagnosis serves as the primary determinant. Users accessing depression-related sub-Reddits are deliberately excluded from the depressive category; instead, they are considered interested parties, potentially seeking information due to proximity to affected individuals. Each entry within the dataset is uniquely identified by Post-TITLE, Post-ID NUMBER, Post-TEXT, and Post-DATE.

The dataset's train-test split allocates 479 subjects to the training set, including 83 positive cases, and 412 subjects to the test set, comprising 54 positive cases. This distribution highlights the importance of a balanced split in machine learning studies, especially for binary classification tasks. A well-considered train-test division ensures the model learns from a representative sample during training while being evaluated on independent data, providing a more accurate assessment of its generalization capabilities. Including positive and negative cases in both sets helps mitigate potential biases. It allows for a comprehensive evaluation of the model's performance across different class distributions, contributing to the study's robustness and reliability.

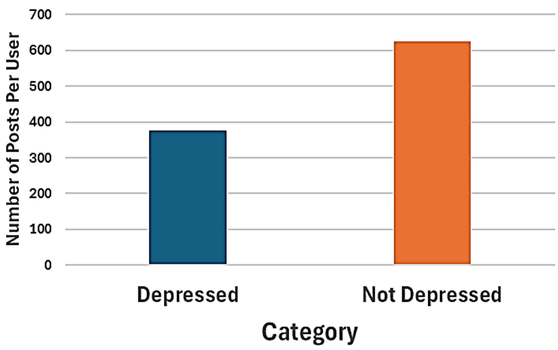


Figure 1. Number of posts made by each user, from those who are and are not depressed

3.2. Dataset preprocessing

Data cleaning precedes feature generation, which is crucial for Reddit datasets comprising diverse subsets. The initial process involved concatenating each user's N most recent titles or sentences; N represents the minimum posts required for model functionality. Authors extracted time intervals between consecutive posts or comments by individual users. Post-grouping procedures included the removal of text-based links, elimination of capitalization, and creation of "bag of words" representations for sequences. The subsequent step utilized WordNetLemmatizer to generate lemmas for each term.

This preprocessing pipeline addresses unique challenges posed by social media data. Concatenation of recent posts ensures sufficient content for analysis, while time interval extraction captures temporal patterns. Removal of links and capitalization standardizes text, reducing noise. "Bag of words" representation transforms text into a format suitable for machine learning algorithms. Lemmatization via WordNetLemmatizer reduces inflected words to base forms, enhancing text normalization and facilitating more accurate linguistic analysis.

The tokenization technique segments the text into discrete words or tokens as part of the lemmatization process. Stop words, such as “and”, “the”, and “is”, were removed to decrease the amount of irrelevant information in the data. In addition, used stemming to normalize the text further by reducing inflected terms to their base form. Subsequently, methodologies such as TF-IDF are used to assess the significance of terms in the dataset. Furthermore, sentiment evaluation is applied to evaluate the psychological nature of the piece of writing, which might provide helpful information for modeling. Addressing any traces of missing data or NaN values is a vital process. The researchers used techniques such as imputation or exclusion, considering the dataset’s specific features. In addition, they verified the reliability and accuracy of the data to guarantee that the created characteristics were dependable for modeling purposes. The data was analyzed using exploratory data analysis (EDA) tools, including visualizations and summary statistics, to provide more insights into the distribution and trends of the data [21].

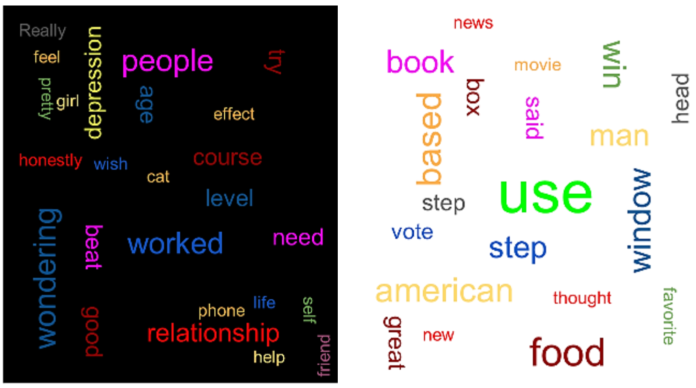
3.3. Models

The study employed multiple models considering the aforementioned factors, comparing against baseline models. Three naive baselines were utilized:

- Random guesser: Fundamental model making arbitrary predictions regarding user depression levels. Equal probability for each estimate.
- Stratified random guesser: Slightly advanced model, estimate not entirely random. Utilizes the percentage of depressed individuals in the training dataset to determine user depression status.
- Logistic regression (LR) TF-IDF models: This model employs the TF-IDF statistic for word vectorization, assessing word importance within the document corpus. The TF-IDF value increases proportionally with word frequency in text, offset by document count containing terms [22]. A logistic regression classifier is applied for the classification task, demonstrating efficacy in various text classification scenarios [23].

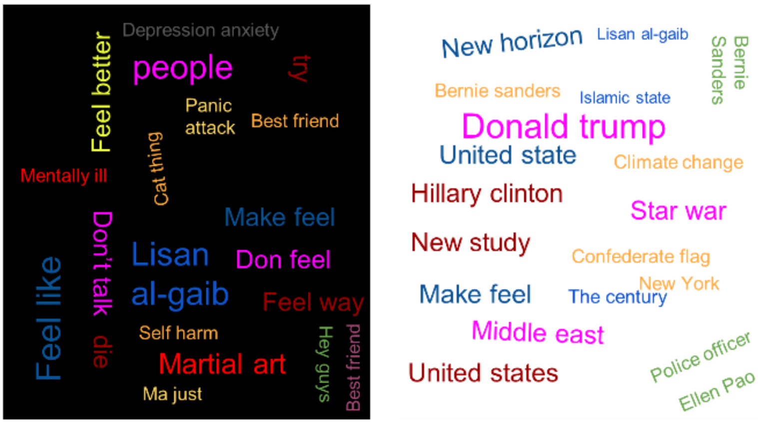
This approach combines statistical text representation (TF-IDF) with probabilistic classification (logistic regression). TF-IDF captures word relevance across the corpus, while logistic regression provides interpretable binary classification. Comparison against naive baselines allows quantification of model improvements over random guessing, establishing performance benchmarks for depression detection tasks. The number of posts (POST LEN), sentiment analysis (SENTIMENT), the average time between posts (AVG DIFF BETWEEN POSTS), and average post length all significantly enhanced our data analysis. These improvements bolstered the LR TF-IDF model, adhering to industry standards. A grid search optimized both the model and TF-IDF vectorizer settings.

Figures 2 and 3 illustrate the key textual features used to classify depression and non-depression based on the post-text. Figure 2 focuses on unigrams, highlighting the most influential single words for each classification: words contributing to the identification of depression in Figures 2(a) and 3(a), and those indicating non-depression in Figures 2(b) and 3(b). Figure 3 extends this analysis to bigrams, showing the pairs of words that play a crucial role in the classification process. The most profound impact on the results was the incorporation of bigram and unigram terms, illustrated in Figures 2 and 3, and filtering out keywords with low document frequency (cut-off value).



(a) (b)

Figure 2. The most influential unigrams in the classification of (a) depression and (b) non-depression



(a) (b)

Figure 3. Bigrams that contribute the most to the classification of (a) depression and (b) non-depression

4. RESULTS AND DISCUSSION
4.1. Results
Table 1 shows the results of various models and their F1 scores. F1 scores were calculated over the seven models' performances in classifying 5, 10, 20, 50, 100, 200, and 500 posts. Calculating F1 scores on several sample sizes allows for assessing how models perform across probabilistic tolerances, revealing their consistency and reliability in increasingly diverse scenarios.

Table 1. F1 results from various models

Model	F1 @ 5	F1 @ 10	F1 @ 20	F1 @ 50	F1 @ 100	F1 @ 200	F1 @ 500
STRATIFIED F1	0.2099	0.2099	0.2099	0.2099	0.2099	0.2099	0.2099
RANDOM F1	0.1921	0.1921	0.1921	0.1921	0.1921	0.1921	0.1921
LR TF-IDF	0.4322	0.5421	0.5901	0.651	0.6981	0.6852	0.6551
LR TF-IDF+SENTIMENT	0.2511	0.4911	0.5412	0.5911	0.6555	0.6312	0.6371
LR TF-IDF+POST LEN	0.3511	0.5001	0.5611	0.6311	0.6711	0.6512	0.6311
LR TF-IDF+POST LEN + SENTIMENT	0.3581	0.5011	0.5711	0.6100	0.6523	0.6411	0.6355
LR TF-IDF AVG DIFF BETWEEN POSTS (Proposed model)	0.4312	0.5177	0.5661	0.6411	0.7011	0.6811	0.6711

4.2. Discussion
Model performance surpasses Losada and Crestani's seminal study regarding F1 scores on initial 10, 100, and 500 posts [6]. The proposed model also outperforms Banovic *et al.'s* F1 scores across all relevant data subsets [19]. Data analysis yielded intriguing insights, as depicted in Table 1. However, average inter-post time interval emerged as the sole feature significantly enhancing basic LR+TF-IDF model performance. Additional features introduced unnecessary noise, diminishing model accuracy on the test set. Figure 4 provides a visual representation of Table 1 results.

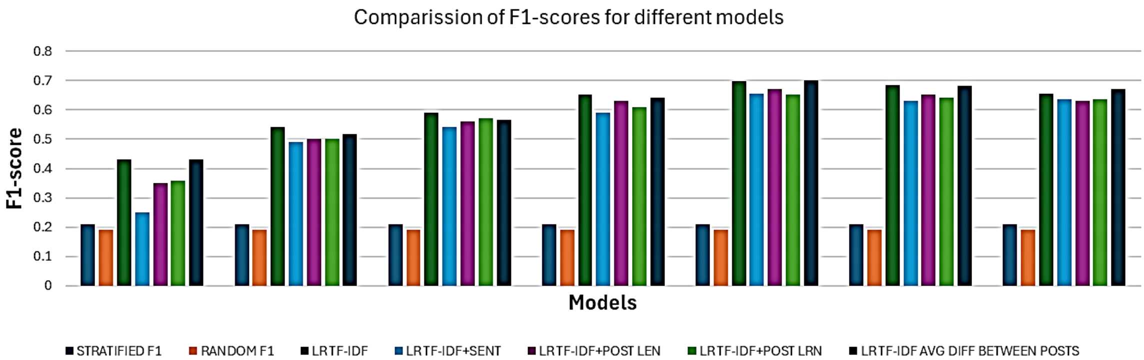


Figure 4. Comparison of F1-scores for different models

The suggested model has demonstrated superior performance compared to its predecessors. Therefore, it is essential to consider the implications that the model's attention weightage can provide to researchers, particularly in identifying trends related to the diagnosis of depression. Utilizing the logistic regression model enables the authors to pinpoint the words and phrases that carry the most weight in influencing a user's classification. To visually represent these influential terms, the authors have employed a word cloud, where the size of each term corresponds to its attention weightage. Analysis of this unigram word cloud clearly indicates specific terms strongly linked to individuals experiencing depression. Consequently, exploring how the model arranges words across different classes becomes pertinent. Additionally, reevaluation of bigram concepts offers insights into their alignment with human intuition. This empirical data reinforces the notion that individuals with depression tend to prioritize self-focused themes over external considerations.

4.3. Future works and implications
4.3.1. Implications

Addressing ethical and privacy concerns in data collection, storage, and utilization is critical [24], necessitating robust anonymization measures. It is essential to engage with the Reddit community for ongoing feedback, continuously monitor user behavior, and adapt to Reddit dynamics. Collaboration with domain experts can significantly enhance evaluation processes, while sensitivity tests contribute to greater transparency. Managing imbalanced data and exploring text preprocessing techniques are pivotal for improving generalization and feature extraction capabilities. Implementing a user-friendly interface or API facilitates seamless integration into diverse applications. The research by Benton *et al.* [25] on gender representation as an additional attribute offers valuable insights.

4.3.2. Future works.

The main challenge in creating the model was the dataset's size and complexity. Expanding the dataset, though costly, will improve performance. Currently, only posting times and users' posts are considered, but adding more user information would benefit academics. Gathering diverse Reddit data from multiple subreddits and demographics can improve understanding of user behavior and predictive potential. Additional variables like age, geography, or hobbies, as well as engagement metrics such as upvotes, downvotes, and comments, would enhance accuracy. Exploring machine learning techniques like neural networks or ensemble approaches, fine-tuning hyperparameters, and using advanced NLP techniques can boost performance. Leveraging long short-term memory (LSTM) networks or utilizing word embeddings for logistic regression holds promise for enhancing performance. Employing explainable AI methodologies, such as feature importance analysis, Shipley additive explanations (SHAP) values, and local interpretable model-agnostic explanations (LIME), enhances transparency and fosters trustworthiness [26]. Effective visualization techniques play a crucial role in intuitively communicating complex concepts, bolstering user confidence and adherence to legal and ethical guidelines.

5. CONCLUSION

Depression, a mental condition, can lead to severe consequences, including suicide, without prompt and appropriate treatment. Comprehensive analysis of posting histories could offer critical insights for improved patient diagnosis, making this research valuable for psychiatrists. The authors utilized TF-IDF and other features to enhance the diagnostic capabilities of the logistic regression-based model. Among these features—post sentiment, count, length, average time between posts—only response time between posts significantly enhanced the model's performance on specific data subsets. Compared to the model proposed by Losada and Crestani, the authors achieved a notably higher F1 score. Methodology highlights the primary objective: identifying words or word groups that aid in depression diagnosis. Employing deep learning models and incorporating additional features such as user gender or experimenting with diverse word embeddings could further enhance the research.

REFERENCES

[1] J. Ma, H. Zhou, Q. Fu, and G. Lu, "Facilitators and barriers in the development and implementation of depression prevention and treatment policies in China: a qualitative study," *BMC Public Health*, vol. 23, no. 1, Feb. 2023, doi: 10.1186/s12889-023-15201-0.

[2] W. Bucci and N. Freedman, "The language of depression," *Bulletin of the Menninger Clinic*. Accessed: Nov. 08, 2022. [Online]. Available: https://www.proquest.com/openview/d804439a2c70467603bbdf0c20a3f31a/1?pq-origsite=gscholar&cbl=1818298

[3] S. Narne, T. Adedoja, M. Mohan, and T. Ayyalasomayajula, "AI-driven decision support systems in management: enhancing strategic planning and execution," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 12, no. 1, pp. 268–276, Mar. 2024.

[4] J. Kim, J. Lee, E. Park, and J. Han, "A deep learning model for detecting mental illness from user content on social media," *Scientific Reports*, vol. 10, no. 1, Jul. 2020, doi: 10.1038/s41598-020-68764-y.

[5] R. Martínez-Castaño, A. Htait, L. Azzopardi, and Y. Moshfeghi, "Early risk detection of self-harm using BERT-based

transformers," in *Studies in Computational Intelligence*, Springer International Publishing, 2022, pp. 183–206. doi: 10.1007/978-3-031-04431-1_8.
[6] D. E. Losada and F. Crestani, "A test collection for research on depression and language use," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Springer International Publishing, 2016, pp. 28–39. doi: 10.1007/978-3-319-44564-9_3.
[7] W. Zaghouani, "A large-scale social media corpus for the detection of youth depression (project note)," *Procedia Computer Science*, vol. 142, pp. 347–351, 2018, doi: 10.1016/j.procs.2018.10.483.
[8] A. Vij and J. Pruthi, "An automated Psychometric analyzer based on sentiment analysis and emotion recognition for healthcare," *Procedia Computer Science*, vol. 132, pp. 1184–1191, 2018, doi: 10.1016/j.procs.2018.05.033.
[9] S. Almouzini, M. Khemakhem, and A. Alageel, "Detecting Arabic depressed users from Twitter data," *Procedia Computer Science*, vol. 163, pp. 257–265, 2019, doi: 10.1016/j.procs.2019.12.107.
[10] A. Priya, S. Garg, and N. P. Tigga, "Predicting anxiety, depression and stress in modern life using machine learning algorithms," *Procedia Computer Science*, vol. 167, pp. 1258–1267, 2020, doi: 10.1016/j.procs.2020.03.442.
[11] J. L. Feuston and A. M. Piper, "Beyond the coded gaze: analyzing expression of mental health and illness on Instagram," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, pp. 1–21, Nov. 2018, doi: 10.1145/3274320.
[12] E. L. Murnane, T. G. Walker, B. Tench, S. Voida, and J. Snyder, "Personal informatics in interpersonal contexts: Towards the design of technology that supports the social ecologies of long-term mental health management," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, pp. 1–27, Nov. 2018, doi: 10.1145/3274396.
[13] J. A. Pater, B. Farrington, A. Brown, L. E. Reining, T. Toscos, and E. D. Mynatt, "Exploring indicators of digital self-harm with eating disorder patients: A case study," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–26, Nov. 2019, doi: 10.1145/3359186.
[14] X. Xu *et al.*, "Leveraging routine behavior and contextually-filtered features for depression detection among college students," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, pp. 1–33, Sep. 2019, doi: 10.1145/3351274.
[15] A. Trifan, R. Antunes, S. Matos, and J. L. Oliveira, "Understanding depression from psycholinguistic patterns in social media texts," in *Advances in Information Retrieval*, Springer International Publishing, 2020, pp. 402–409. doi: 10.1007/978-3-030-45442-5_50.
[16] P. Mathur, R. Sawhney, S. Chopra, M. Leekha, and R. Ratn Shah, "Utilizing temporal psycholinguistic cues for suicidal intent estimation," in *Advances in Information Retrieval*, Springer International Publishing, 2020, pp. 265–271. doi: 10.1007/978-3-030-45442-5_33.
[17] X. Wang, C. Zhang, Y. Ji, L. Sun, L. Wu, and Z. Bao, "A depression detection model based on sentiment analysis in micro-blog social network," in *Trends and Applications in Knowledge Discovery and Data Mining*, Springer Berlin Heidelberg, 2013, pp. 201–213. doi: 10.1007/978-3-642-40319-4_18.
[18] M. Song *et al.*, "Daily mental health monitoring from speech: A real-world Japanese dataset and multitask learning analysis," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Jun. 2023. doi: 10.1109/icassp49357.2023.10096884.
[19] L. Banović, V. Fatorić, and D. Rakovac, "How soon can we detect depression?" Text Analysis and Retrieval 2019 Course Project Reports, [Online]. Available: https://www.fer.unizg.hr/_download/repository/TAR-2019-ProjectReports.pdf#page=7
[20] M. Al-Mosaiwi and T. Johnstone, "In an absolute state: elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation," *Clinical Psychological Science*, vol. 6, no. 4, pp. 529–542, Jan. 2018, doi: 10.1177/2167702617747074.
[21] O. Berjawi, G. Fenza, and V. Loia, "A comprehensive survey of detection and prevention approaches for online radicalization: Identifying gaps and future directions," *IEEE Access*, vol. 11, pp. 120463–120491, 2023, doi: 10.1109/access.2023.3326995.
[22] R. K. Dey and A. K. Das, "Modified term frequency-inverse document frequency based deep hybrid framework for sentiment analysis," *Multimedia Tools and Applications*, vol. 82, no. 21, pp. 32967–32990, Mar. 2023, doi: 10.1007/s11042-023-14653-1.
[23] K. Shah, H. Patel, D. Sanghvi, and M. Shah, "A comparative analysis of logistic regression, random forest and KNN models for the text classification," *Augmented Human Research*, vol. 5, no. 1, Mar. 2020, doi: 10.1007/s41133-020-00032-0.
[24] S. Dodda, A. Kumar, N. Kamuni, and M. M. T. Ayyalasomayajula, "Exploring strategies for privacy-preserving machine learning in distributed environments," *3rd IEEE International Conference on Artificial Intelligence for Internet of Things AIIoT 2024*, Vellore, India, Apr. 2024, doi: 10.1109/AIIoT58432.2024.10574614.
[25] A. Benton, M. Mitchell, and D. Hovy, "Multitask learning for mental health conditions with limited social media data," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Association for Computational Linguistics, 2017, doi: 10.18653/v1/e17-1015.
[26] M. M. T. Ayyalasomayajula, S. Ayyalasomayajula, and J. K. Pandey, "Explainable artificial intelligence (XAI) for emotion detection," in *Machine and Deep Learning Techniques for Emotion Detection*, IGI Global, 2024, pp. 203–232, doi: 10.4018/979-8-3693-4143-8.ch010.

BIOGRAPHIES OF AUTHORS

Madan Mohan Tito Ayyalasomayajula is a seasoned researcher and author with over two decades of expertise in scalable data architecture, distributed computing, cloud computing, machine learning, and artificial intelligence. He holds a doctorate in computer science from Aspen University, complemented by a master of computer applications and a master of technology in computer science and engineering from Osmania University. Throughout his career, Madan has collaborated closely with global organizations, specializing in architecting and implementing scalable data management solutions for big data, ML, and AI. He has contributed significantly to Mahindra Comviva, developing products for mobile network operators worldwide, and has provided consulting services to Toyota Motor North America and Walgreens. Currently serving as a Senior Technology Architect at Infosys, Madan continues to innovate in data solutions, including ML and AI, for a diverse client base. He can be contacted at mail2tito@gmail.com.

Akshay Agarwal received bachelor of technology degree in computer science and engineering from UPTU, India in 2009 and the master's degree in computer science from Illinois Institute of Technology, Chicago, USA in 2012. He is a seasoned data engineer and architect with over fourteen years of software industry experience skilled in architecture, designing, developing and implementing IT business solutions. He is dedicated to staying abreast of emerging trends and best practices in the rapidly evolving field of data engineering to drive innovation to meet overarching business goals. He initially worked with Infosys in the data transformation space and then worked with Capital One in the credit card, risk management, digital, and audit and compliance lines of business. Currently, with Fourth Enterprises spearheading their data analytics, artificial intelligence, reporting, and cloud migration initiatives. He is associated with the British Computer Society with current research interests in data analytics, AI/ML and natural language processing He can be contacted at akshaydata.ai@gmail.com.

Shahnawaz Khan is a seasoned IT professional with over 18 years of experience specializing in product innovation management and robotic process automation (RPA). Shahnawaz completed master of computer application from Bundelkhand University Jhansi, UP India. Currently he is Based in Surprise, Arizona, Shahnawaz has honed his skills in various domains including SDLC, STLC, big data, AI and ML, insights reporting, and RPA. With a track record of managing complex projects and eliciting requirements, Shahnawaz has led end-to-end projects in industries such as travel, insurance, and procurement. His expertise spans data analysis, data science, visualization, AI and ML, and automation, with proficiency in tools like Birst, VSTS, and Selenium. He can be contacted at shanawaz.qa@gmail.com.