Journal of Advances in Developmental Research (IJAIDR)
E-ISSN: 0976-4844 ● Website: www.ijaidr.com ● Email: editor@ijaidr.com

Big Data and Mental Health: Social Media Sentiment Analysis

Srinivasa Kalyan Vangibhurathachhi

Srinivasa.Kalyan2627@gmail.com

Abstract:
With mental health problems affecting 1 in every 8 people globally, leveraging technologies such as big data and sentimental analysis techniques for detection and monitoring is crucial. This paper explores how big data and sentiment analysis techniques can be used to predict mental health issues from posts of social media users. The research finding shows that sentimental analysis is an innovative solution for detecting and monitoring mental health issues among social media users through the collected big data sets. Key findings indicate that lexicon-based approaches such as VADER and machine learning models like SVM and Naïve Bayes present potential advantages and limitations in identifying and monitoring mental health issues. Also, hybrid approaches that synergises machine learning methods and lexicon-based approaches presents potential benefits in addressing mental health concerns among social media users. Current advances in sentimental analysis techniques for identifying and monitoring mental health issues from various data sources are also presented in the paper. The research also identified challenges such as inherent subjectivity in labelling, interpretational complexities, data quality and privacy concerns. Looking forward, future researchers should explore multilingual sentiment analysis, cross-platform analysis, integrate wearables and mobile apps and conduct longitudinal studies. In summary, while sentimental analysis presents promising avenue for mental health detection, there is need for continuous refinement and ethical adherence.

Keywords: mental health, social media, big data, sentiment analysis, machine learning algorithms, hybrid models, lexicon-based methods.

1. INTRODUCTION

Globally, the prevalence of mental health problems is at an all-time high with the World Health Organization (2022) reporting that 1 in every 8 people live with mental disorders such as depression, anxiety, bipolar disorder, post-traumatic disorder, schizophrenia among others. With the growing use of social media platforms such as Facebook, X, Tik Tok, YouTube among others, users to tend express their emotions, thoughts, feelings and experiences related to mental health (Babu & Kanaga, 2022). In particular, millions of people express their moods, frustrations or even daily mental health struggles online which creates a rich source of big data that can be analysed for patterns and trends on mental health. As a result, big data technologies such as sentiment analysis which monitors and addresses mental health issues has become a crucial area of research. As a branch of NLP, sentimental analysis involves identification, extraction and interpretation of opinions, emotions and attitudes expressed in textual data (Bharadwaj et al., 2025). In mental health, sentimental analysis has the potential to offer insights into sentimental shifts, emotional states and indicators of mental health problems. Obagbuwa et al. (2023) affirms that social media sentiment analysis enables real-time monitoring of a large populations' emotional states thus potentially predicting mental health issues before they escalate. To this end, this paper explores how big data and sentiment analysis techniques can be used to predict mental health issues from posts of social media users by focussing on the following specific objectives;

i) Examining capabilities sentimental analysis techniques in detecting mental health trends among social media users

ii) Assessing current advancements in mental health problem detection and sentimental analysis
iii) Identifying the impact of sentiment analysis of social media data on early intervention, prevention and delivery of mental health care
iv) Identifying challenges and research directions related to big data and sentimental analysis addressing mental health problems among social media users.

2. PROBLEM STATEMENT:

While mental health issues are a global crisis, many affected individuals do not seek help due to stigma, lack of access to healthcare or lack of awareness about mental health problems. Conway and O'Connor (2017) highlight that traditional techniques of diagnosis tend to fail in detecting mental health issues in real-time especially when individuals who may not present symptoms during clinical consultations. Yuan (2024) adds that the social stigma associated with mental health disorders may prevent individuals from seeking help and during consultations, many individuals may not present their full mental health concerns. In light of this, social media sentiment analysis has been explored for monitoring the emotional states of large populations in real time from various social media platforms, thus enabling the early identification of mental health issues before they escalate (Obagbuwa et al., 2023).

3.0 PROPOSED SOLUTION:

3.1 Social media sentimental analysis as a solution to address mental health problems

By definition, sentimental analysis involves identifying, extracting and organising sentiments from user generated texts, visual or audio data from social platforms (Babu & Kanaga, 2022). Through sentiment analysis, machine learning models can assess and identify social media posts related to mental health issues. Benrouba and Boudour (2023) argues that sentiment analysis models can classify texts into positive, negative or neutral sentiments which reveals patterns on mental health discussions. By applying sentiment analysis to textual data from platforms like Twitter or Facebook, subtle shifts in users' emotional states can be identified. Early identification of these patterns could lead to earlier interventions and more personalized mental health support (Conway & O'Connor, 2017). For instance, recognizing negative sentiment trends in posts could signal the onset of depression or anxiety, allowing for timely mental health outreach.

3.1.1. Sentimental Analysis Methods for Mental Health monitoring

Mental health monitoring through sentimental analysis entails extracting emotional cues from textual data thus enabling identification of individuals at risk. Bharadwaj et al. (2025) highlighted three key techniques for monitoring mental health issues such as machine learning algorithm, lexicon-based approach and hybrid approach as shown in figure 1 below. Each of these methods provides specific advantages and challenges in mental health related sentiment patterns within social media platforms.

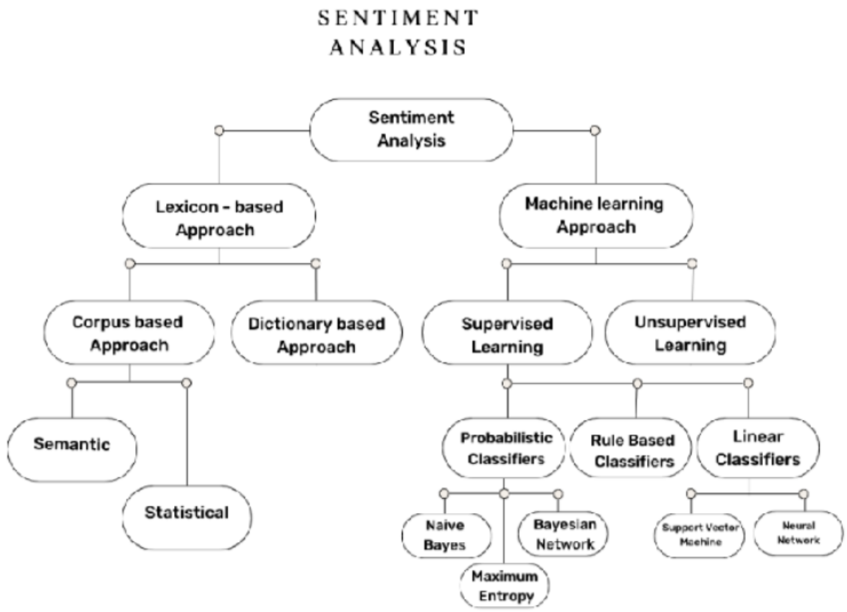

Figure 1: Flow chart of sentimental analysis (Bharadwaj et al., 2025)

Starting with lexicon-based approach, the technique offers initial method of gauging sentiment in textual data from social media platforms by identifying and assessing mental health contents based on predefined sentiment lexicons that associate words with sentiment labels (Bonta et al., 2019). Although simplistic, this method offers valuable insights on the emotional state of individuals using online platforms. An example of lexicon-based approach is the Valence Aware Dictionary and Sentiment Reasoner (VADER) which quantifies sentiments from a piece of text into a single numerical score (Bonta et al., 2019). The VADER model uses dictionary of words lexicon that has predefined scores which represents positive or negative terms and another set of rules to evaluate sentimental texts. Thereafter, scores are combined using specific formula to generate composite score. While VADER can evaluate short individual words and idiomatic phrases to capture contexts and emotional intensity within texts, Zahoor and Rohila (2020) observed that its performance may vary in other domains. Accordingly, it might not handle complex sentences or sarcasm effectively in social media platforms. Bharadwaj et al. (2025) suggested the following mathematical formula for calculating compound score from the sum of valence scores of each word within textual data, normalized between -1 and 1.

$$C = \frac{\sum \text{valence scores}}{N \times \text{maximum valence scores}},$$

where C = The compound score, valence scores = Individual valence scores of words, N = Total number of words in the text, and maximum valence score = the maximum possible valence score.

Another example of lexicon based approach is the sent WordNet and Text Blob which assigns pre-calculated sentiment scores to words and calculates sentiment score for textual data based on the word scores (Bonta et al., 2019). While sent WordNet assigns scores in line with word meanings from the word net database, text Blob uses predefined set scores for words to evaluate the sentiment. Menaha and Ananthi (2024) argued that both models can quickly gauge sentiment from text by looking at pre-assigned scores of words within a particular corpus. However, the models lack ability to capture contexts effectively because they struggle with nuanced text expressions. Unlike VADER, these methods do not have single mathematical formula for sentiment score calculation. In light of these discussions, Bharadwaj et al. (2025) assert that lexicon-based approach's speed and simplicity qualifies it for quick sentiment analysis of large datasets on mental health issues. However, the performance could be affected by use of nuanced language, contextual complexities and sarcasm. Al-Qablan et al. (2023) adds that lexicon based approaches offers rapid insights into sentiment patterns, provides transparency which allows users to easily comprehend sentiments. However, the presence of words with double meanings (positive and negative sentiments) may result in misclassifications. Al-Qablan et al. (2023) suggested a flowchart as shown in figure two below.

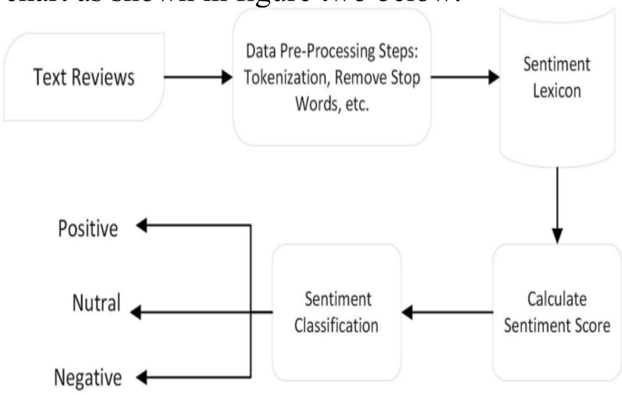

Figure 2: Flow chart for lexicon-based approaches for sentiment analysis (Al-Qablan et al., 2023)

Looking at machine learning algorithms for mental health monitoring, research indicates that it it's a potent framework that leverages on data-driven insights. Rahat et al. (2019) argues that machine learning algorithms can learn patterns from data sets to discern critical nuances within sentiment expressions so as to gain deeper understanding of emotional states in social media platforms. Existing literatures discuss a number of machine learning models used in sentiment analysis. A popular machine learning model for mental health monitoring is the Naive Bayes which computes the probability of specific class given the input features (Rahat et al., 2019). In the context of mental health monitoring, data cleaning and pre-processing involves removing extraneous elements from textual data while tokenising it into individual units like words or n-grams (Goel et al., 2016). When applying Naive Bayes, the data is changed to numerical feature vectors using techniques like TF-IDF or BoW that capture word frequencies. During training, the algorithm computes probabilities in line with occurrence of labelled dataset on mental health. The trained Naive Bayes classifier can predict the sentiment label based on computed probabilities by selecting the class with the higher probability. While Naïve Bayes is simple, efficient and effectively works with small amount of data, Goel et al. (2016) observed that it heavily relies on assumptions of independence between features.

As a machine learning algorithm, Support vector machines (SVM) have also been proposed for classifying sentiment analysis within mental health monitoring. Similar to Naïve Bayes, SVM feature extraction focuses on converting textual data to numerical feature using techniques like TF-IDF or BoW (Rahat et al., 2019). The algorithm is then trained to maximally separate two dataset classes into either mental health disorder or non-mental health disorder. Accordingly, the algorithm can predict sentiment label by evaluating position of input text in relation to learned hyperplane. While SVM is effective with high dimensional data and handles non-linear relationship, Alam and Yao (2019) noted that it calls for extensive training for optimal performance. Overall, incorporating SVM and Naïve Bayes into mental health monitoring can harness complex sentiment expressions from social media platforms thus contributing to identification and intervention of mental health issues. Figure three below illustrates the flow chart on machine learning approach to sentiment analysis

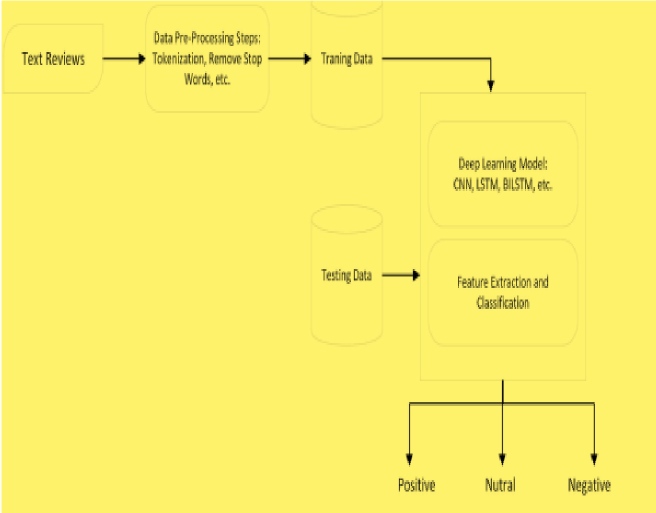

Figure 3; Machine learning based approach to sentiment analysis (Al-Qablan et al., 2023)

Finally, the hybrid approach for monitoring mental health integrates the strengths of lexicon-based and machine learning techniques to develop synergistic framework for sentiment analysis (Bharadwaj et al., 2025). For lexicon-based analysis, textual data from social media platforms undergoes sentiment analysis using lexicon techniques like VADER to assess emotional tone of contents. On its part, the machine learning integration obtain sentiment scores from lexicon-based analysis and other textual features. Al-Qablan et al. (2023) point out that hybrid approach is advantageous because it leverages interpretability of lexicon-based SA and harnesses the predictive power of machine learning algorithm. Notably, integrating these techniques helps in getting a nuanced understanding of sentiment expression which helps in identifying individuals at risk of mental health issues. However, integration challenges exist when lexicon sentiment assessment does not align with true sentiment (Bharadwaj et al., 2025). Figure 4 below highlights the flowchart for hybrid based model in sentiment analysis. Table 1 below comparatively evaluates sentiment analysis approaches.

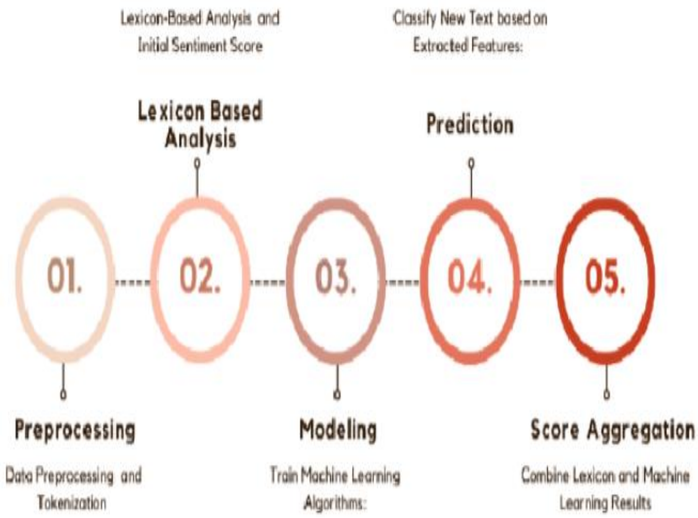

Figure 4; flow chart of hybrid based approach to sentiment analysis (Bharadwaj et al., 2025)

	Lexicon-based approach	Machine learning-based approach
Type	Dictionary and corpus-based	Supervised, Sime-supervised, and Unsupervised
Labelled Data	No need	Supervised and semi-supervised approaches required labelled data
Feature extraction	Manual feature engineering	Automated feature extraction
Time to train the data	No need	Need time
Complexity	Simple and fast	Depending on the algorithm employed
Domain dependency	Lexicons are usually designed for specific domains	Domain dependence, but they are often more malleable and adaptable than lexicon-based methods
Adaptability	Requires adding new words and terms to lexicons for new domains	Flexibly adapt to a wide range of domains with adequate training
Performance	Depends on the lexicon scope	Depends on the quality of labelled data
Applications	Simple SA tasks	Various SA tasks
Out-of-vocabulary words	May have trouble with words that are not in the lexicon	Can to a limited degree handle out-of-vocabulary words by utilizing n-gram representations
Amount of training data	It is dependent on pre-defined lexicons	Requires an adequate quantity of labelled training data

Table 1: comparative evaluation of sentiment analysis approaches (Al-Qablan et al., 2023)

3.1.2 current advancements in Mental Health problem detection and sentimental analysis

There is a growing body of research on mental health problem detection using sentimental analysis techniques. Huang et al (2021) deployed convolutional long short term memory (LSTM) model to attain high sentiment categorisation performance among social media users exhibiting attention-emotion behaviours. They found reduced performance and robustness due to the inability of the model to balance ideal parameters. A research by Gandhi et al. (2021) employed LTSM and convolutional neural networks models to identify frequently occurring mental health terms on Twitter platform. They generated huge textual data, identified individuals who could be affected by mental health issues and recognized the need for improvements in feature learning methods. Similar to Gandhi et al. (2021), Jain et al. (2021) utilised CNN-LSTM model for unstructured user review analysis by focusing specifically on data from social media platforms and found positive relationship between the models effectiveness and mental health disorders among users. Further, Kaur et al. (2021) suggested Hybrid Heterogeneous Support Vector Machine (H-SVM) approach which extracts tweets from users of several languages. Their data collection was limited to Twitter and traditional vector representation which reduced generalisability of findings.

Recent research by Suman et al. (2020) created a new sentiment analysis engine for detecting depression status among social media users. The study offered a unique perspective for estimating depression from the initial state to the final stages. Jo et al. (2017) investigated the relationship between computational models and psychological states of social media users and determined that sentiment analysis can identify mental health issues and help in suggesting appropriate intervention methods based. Finally, Li et al. (2018) explored the impact of linguistic features in detecting depression stigma on social media users and found that language patterns play a crucial role in detecting social stigma related to depression. Overall, these studies help in advancing sentimental analysis techniques for identifying and monitoring mental health symptoms from various data sources.

4.0 IMPACT OF SENTIMENT ANALYSIS OF SOCIAL MEDIA DATA ON EARLY INTERVENTION, PREVENTION AND DELIVERY OF MENTAL HEALTH CARE

Existing studies have shown that the impact of big data and sentiment analysis on mental health care is profound. By detecting negative sentiment and emotions, Gandhi et al. (2021) determined that sentiment analysis could allow for early detection of mental health issues and promote timely interventions. Al-Qablan et al. (2023) added that data-driven insights can help mental health professionals to personalise treatment plans to individual needs. More important, Babu and Kanaga (2021) assert that sentiment analysis can be used in public health campaigns in raising awareness about mental health thus combating stigma and encouraging individuals to seek help.

5.0 CHALLENGES AND FUTURE RESEARCH DIRECTIONS

5.1 Challenges

Bharadwaj et al. (2025) opines that the inherent subjectivity when labelling mental health disorder from text presents significant challenge. Precisely, it is challenging when defining and labelling a text as mental health disorder or non-mental health disorder because of several emotional expressions and sensitivity of contexts. Additionally, analysing personal textual data when monitoring mental health raises privacy concerns (Obagbuwa et al., 2023). This calls for ethical considerations such as stringent anonymization and privacy protection when handling sensitive data. Tušl et al. (2022) adds that getting high quality and diverse labelled datasets for training models remains a key challenge. The limited nature of labelled and standardised datasets that encompass a wide range of cultures, demographics and linguistics affects the effectiveness of sentiment analysis. As previously discussed, sarcasm and context in texts present challenges in sentiment analysis (Bharadwaj et al., 2025). Accordingly, understanding sarcasm and nuanced language which is prevalent in social media forums calls for more sophisticated models to curb misinterpretation.

5.2 future research directions

While the current research highlights the significance of sentiment analysis in mental health research, future researchers can explore several areas. Given that social media data is global and mental health issues vary across languages and cultures, future research should expand sentiment analysis tools to include multiple languages and dialects to gain more inclusive understanding of global mental health issues (Bharadwaj et al., 2025). At the moment, majority of research reviewed above focuses on single platforms like Twitter or Facebook. Future researchers should conduct cross-platform analysis which incorporates data from Facebook, Reddit, Instagram and other platforms to gain more comprehensive understanding of mental health trends. Also, the integration of sentiment analysis with data from mental health apps and wearable devices could offer even more precise monitoring of emotional and mental well-being (Jain et al., 2021). This will provide a holistic approach to mental health care. Further, longitudinal studies that track sentiment over time can offer insights on how mental health issues evolves and thus can predict the onset of disorders before they become worse.

6.0 CONCLUSION

The research extensively delves into big data and sentimental analysis as innovative solution for detecting and monitoring mental health issues through social media platforms. the research outlines various methods starting with lexicon-based approaches such as VADER to machine learning models like SVM and Naïve Bayes. Also, the research presents hybrid approaches that synergises machine learning methods and lexicon-based approaches. Current advances in sentimental analysis techniques for identifying and monitoring mental health issues from various data sources are also presented. While exploring the research area, the paper identified challenges such as inherent subjectivity in labelling, interpretational complexities, data quality and privacy concerns. Looking forward, future researchers should explore multilingual sentiment analysis, cross-platform analysis, integrate wearables and mobile apps and conduct longitudinal studies. In summary, while sentimental analysis presents promising avenue for mental health detection and monitoring, there is need for continuous refinement and ethical adherence.

REFERENCES:

1. A. Goel, J. Gautam and S. Kumar, "Real time sentiment analysis of tweets using Naive Bayes," *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, Dehradun, India, pp. 257-261, 2016, doi: 10.1109/NGCT.2016.7877424.
2. A. Li, D. Jiao, and T. Zhu, "Detecting depression stigma on social media: A linguistic analysis," *Journal of Affective Disorders*, vol. 232, pp. 358–362, May 2018. doi:10.1016/j.jad.2018.02.087
3. A. M. Rahat, A. Kahir and A. K. M. Masum, "Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset," *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, Moradabad, India, pp. 266-270, 2019, doi: 10.1109/SMART46866.2019.9117512.
4. B. Bharadwaj, S. Nayak and P.K. Panigrahi, "Sentiment Analysis for Identifying Depression through Social Media Texts Using Machine Learning Technique", Big Data and Computing Visions, vol. 5, no. 2, pp.102-118, 2025.
5. F. Benrouba and R. Boudour, "Emotional sentiment analysis of social media content for mental health safety," *Soc. Netw. Anal. Min.* vol. 13, no. 17, 2023, https://doi.org/10.1007/s13278-022-01000-9
6. F. Huang, X. Li, C. Yuan, S. Zhang, J. Zhang, and S. Qiao, "Attention-emotion-enhanced convolutional LSTM for sentiment analysis," *IEEE transactions on neural networks and learning systems,* vol. 33, no. 9, pp. 4332–4345. 2021, https://doi.org/10.1109/TNNLS.2021.3056664
7. H. Kaur, S.U. Ahsaan, B. Alankar and V. Chang, "A proposed sentiment analysis deep learning algorithm for analysing COVID-19 tweets," *Information Systems Frontiers,* vol. 23, no. 6, pp. 1417–1429, 2021, https://doi.org/10.1007/s10796-021-10135-7
8. H. Jo, S.M. Kim and J. Ryu, "What we really want to find by sentiment analysis: the relationship between computational models and psychological state," *arXiv preprint arXiv:1704.03407*, 2017.
9. I.C. Obagbuwa, S. Danster, and O. C. Chibaya, "Supervised machine learning models for depression sentiment analysis," *Frontiers in Artificial Intelligence*, 6(3), pp. 32-49, 2023. doi:10.3389/frai.2023.1230649
10. M. Conway and D. O'Connor, "Social Media, Big Data, and Mental Health: Current advances and ethical implications," *Current Opinion in Psychology*, vol. 9, pp. 77–82, 2016. doi: 10.1016/j.copsyc.2016.01.004
11. M. Tušl, A. Thelen, K. Marcus, A. Peters, E. Shalaeva, B. Scheckel, M. Sykora... and O. Gruebner, "Opportunities and challenges of using social media big data to assess mental health consequences of the COVID-19 crisis and future major events," *Discover Mental Health*, vol. 2, no. 1, pp. 14-32, 2022. doi:10.1007/s44192-022-00017-y
12. N.V. Babu and E.G.M., Kanaga, "Sentiment Analysis in Social Media Data for Depression Detection Using Artificial Intelligence: A Review," *SN COMPUT. SCI, vol.* 3, no. 74, 2022. https://doi.org/10.1007/s42979-021-00958-1
13. P.K. Jain, V. Saravanan and R. Pamula, "A hybrid CNN-LSTM: A deep learning approach for consumer sentiment analysis using qualitative user-generated contents," *Transactions on Asian and low resource language information processing*, vol. 20, no. 5, pp. 1–15, 2021, https://doi.org/10.1145/3457206
14. R. Menaha and K. Ananthi, "Reviewing the effectiveness of lexicon-based techniques for sentiment analysis in Massive open online courses," *International Journal of Data Science and Analytics*, 2024. doi:10.1007/s41060-024-00585-y
15. S.K., Suman, H. Shalu, L.A. Agrawal, A. Agrawal and J. Kadiwala, "A novel sentiment analysis engine for preliminary depression status estimation on social media', https://doi.org/10.48550/arXiv.2011.14280, 2020
16. S. Zahoor and R. Rohilla, "Twitter Sentiment Analysis Using Lexical or Rule Based Approach: A Case Study," *2020 8th International Conference on Reliability, Infocom Technologies and*

Optimization (Trends and Future Directions) (ICRITO), Noida, India, pp. 537-542, 2020. doi: 10.1109/ICRITO48877.2020.9197910.

17. U.D. Gandhi, P. Malarvizhi Kumar, G. Chandra Babu and G. Karthick, "Sentiment analysis on twitter data by using convolutional neural network (CNN) and long short term memory (LSTM)', *Wireless personal communications,* pp. 1–10. https://doi.org/10.1007/s11277-021-08580-3, 2021.
18. V. Bonta, N. Kumaresh, and N. Janardhan, "A comprehensive study on lexicon based approaches for sentiment analysis," *Asian Journal of Computer Science and Technology*, vol. 8, no. S2, pp. 1–6, 2019. doi:10.51983/ajcst-2019.8.s2.2037
19. WHO. *"Mental Disorders: Key Facts."* https://www.who.int/news-room/fact-sheets/detail/mental-disorders (Accessed: 6th May 2025)
20. X. Yuan, "Social Media Sentiment Analysis and Mental Health Prediction using Deep Belief Network (DBN)," *2024 Second International Conference on Networks, Multimedia and Information Technology (NMITCON)*, Bengaluru, India, 2024, pp. 1-5, Doi: 10.1109/NMITCON62075.2024.10699157.