International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

PREDICTIVE MODELS FOR MENTAL HEALTH DETECTION: A COMPARATIVE ANALYSIS OF LOGISTIC REGRESSION, RANDOM FOREST, AND SVM

Shubhi Verma[a], Priyanshi Tyagi[b], Kirti Sharma[c] , Jatin Tyagi[d]

[1]Assistant Professor, Department of Computer Science and Engineering, Shri Ram Group of Colleges, Muzaffarnagar, Uttar Pradesh, India
shubhiverma0007@gmail.com
[2]B.Tech Scholar, Department of Computer Science and Engineering, Shri Ram Group of Colleges, Muzaffarnagar, Uttar Pradesh, India
priyanshityagi2004@gmail.com
[3]B.Tech Scholar, Department of Computer Science and Engineering, Shri Ram Group of Colleges, Muzaffarnagar, Uttar Pradesh, India
kirtishamli@gmail.com
[4]B.Tech Scholar, Department of Computer Science and Engineering, Shri Ram Group of Colleges, Muzaffarnagar, Uttar Pradesh, India
tyagijatin093@gmail.com

ABSTRACT:

Mental health is vital to overall well-being, influencing emotional, psychological, and social aspects of life. It plays a crucial role in determining how individuals handle stress, relate to others, and make decisions. Good mental health is essential for personal and community development, as it affects productivity, relationships, and overall quality of life. This study examines the use of text analysis for early detection of mental health disorders via a dedicated assessment website. We applied natural language processing (NLP) and machine learning techniques, specifically Logistic Regression, Random Forest, and Support Vector Machine (SVM), to classify conditions like borderline personality disorder (BPD), anxiety, depression, bipolar disorder, and schizophrenia. Using a dataset of 35,000 training and 6,108 test records, we tested various models. Logistic Regression achieved the highest accuracy at 91.67%, outperforming Random Forest (79.17%) and Support Vector Machine (SVM) (83.33%). The study demonstrates the potential of these technologies to aid mental health professionals in timely and accurate diagnostics, emphasizing the importance of integrating advanced computational methods like NLP and machine learning algorithms into mental health assessments. Effective early detection and intervention can significantly improve outcomes for individuals suffering from mental health issues, highlighting the critical need for innovative tools and approaches in this field.

Keywords: Mental health, Text analysis, Machine learning, Early detection, NLP

1. INTRODUCTION :

Mental health disorders are defined as syndromes characterized by clinically significant disturbances in an individual's cognition, emotion regulation, or behavior, reflecting dysfunctions in psychological, biological, or developmental processes (American Psychiatric Association, 2013). These disorders have a profound impact on a substantial portion of the global population, with over 1 billion individuals affected, accounting for approximately 16% of the global population. The COVID-19 pandemic has further exacerbated the incidence of mental health issues, underscoring the urgent need for effective assessment and early detection methods. Platforms such as Mind Matters, dedicated to mental health assessment, play a crucial role in identifying and addressing these disorders.

Mind Matters utilizes advanced technology to offer a comprehensive mental health assessment platform that helps users understand their mental health status through interactive questionnaires and detailed analysis. The platform guides users through a series of questions designed to evaluate various aspects of their mental and emotional well-being. Based on the responses, the platform employs sophisticated algorithms to analyze the user's mental state and provides a diagnostic score that helps identify potential mental health issues. This score can serve as a preliminary indication of conditions such as borderline personality disorder (BPD), anxiety, depression, bipolar disorder, schizophrenia, and substance abuse.

Mind Matters goes beyond simply diagnosing mental health issues; it also provides personalized recommendations and resources tailored to each user's specific needs. By analyzing patterns in user responses and behavior, the platform offers insights into potential triggers and coping mechanisms, empowering users to take proactive steps towards better mental well-being. Additionally, Mind Matters fosters a supportive online community where users can connect with others facing similar challenges, fostering a sense of belonging and reducing feelings of isolation.

The platform's user-friendly interface and accessible language make it suitable for individuals of all ages and backgrounds, democratizing access to mental health resources. Through continuous feedback and updates, Mind Matters remains responsive to the evolving needs of its users and the latest developments in mental health research and technology. By combining cutting-edge technology with compassionate support, Mind Matters aims to revolutionize the way mental health is understood, assessed, and addressed in today's digital age.

1.1. Mental Illness Detection

Research on detecting mental illnesses in social media platforms involves complex methodologies and ethical considerations. While social media data can provide valuable insights into individuals' mental health, it's crucial to ensure the privacy and ethical treatment of users' data. Techniques such as web mining and emotion analysis are utilized as preliminary steps to raise awareness and identify potential mental health issues.

Ethical considerations are paramount in handling social media data, with researchers taking steps to anonymize and protect user privacy. However, challenges remain in maintaining anonymity, and researchers must adhere to confidentiality agreements to safeguard user data.

The naturalistic setting of social media interactions offers a rich source of longitudinal data and self-disclosure, which can be valuable for detecting mental illnesses. However, determining the extent of personal disclosure and the accuracy of information is essential for reliable detection.

Feature engineering plays a crucial role in detecting mental illnesses in social media content. The Linguistic Inquiry Word Count (LIWC) lexicon is widely used for extracting lexical features related to various psychological constructs. These features are then analyzed to identify patterns indicative of specific mental disorders.

Despite the progress in feature engineering, challenges remain in identifying unique features for each mental disorder. Researchers strive to extract features that overlap with each other while being distinct to a particular disorder, enhancing the accuracy of detection algorithms.

Research in the field of mental illness detection in social media platforms is multifaceted and dynamic, continually evolving to address new challenges and opportunities. Ethical considerations surrounding the use of social media data underscore the importance of transparency, informed consent, and privacy protection. While anonymization techniques are employed to safeguard user privacy, ongoing efforts are required to mitigate the risk of data breaches and unauthorized access.

The longitudinal nature of social media data offers researchers a unique opportunity to track changes in individuals' mental health over time. This wealth of data enables the exploration of temporal patterns and trends, providing valuable insights into the progression and manifestation of mental illnesses.

In addition to textual data, researchers also leverage multimodal data sources, such as images, videos, and audio recordings, to enrich the analysis of mental health-related content. Integrating these diverse data modalities enhances the depth and granularity of mental illness detection algorithms, allowing for a more comprehensive understanding of individuals' psychological states.

Machine learning and deep learning techniques play a pivotal role in analyzing large-scale social media datasets and extracting meaningful patterns and associations. These advanced algorithms are capable of processing vast amounts of data efficiently, enabling the identification of subtle indicators of mental illness that may not be apparent through manual inspection.

Collaborations between researchers, mental health professionals, and technology experts are essential for advancing the field of mental illness detection in social media. By combining domain expertise with cutting-edge computational methods, interdisciplinary teams can develop innovative solutions that address the complex challenges inherent in mental health assessment.

Furthermore, the integration of user feedback and iterative refinement processes is crucial for enhancing the accuracy and usability of mental illness detection systems. Continuous evaluation and validation of algorithms against ground truth labels ensure the reliability and effectiveness of these systems in real-world scenarios.

Ultimately, the goal of mental illness detection in social media is not only to identify individuals at risk but also to provide timely interventions and support. By leveraging the power of technology and data-driven insights, we can empower individuals to seek help, access resources, and ultimately improve their mental well-being.

However, it is essential to acknowledge the limitations and ethical considerations inherent in this field. Privacy concerns, data biases, and algorithmic transparency are critical areas of concern that must be addressed to ensure the responsible use of social media data for mental health purposes.

In conclusion, mental illness detection in social media holds immense promise for improving our understanding of mental health and enhancing early intervention efforts. Through interdisciplinary collaboration, ethical practice, and ongoing research, we can harness the potential of technology to make meaningful strides in mental health care and support.

2. LITERATURE REVIEW

Sno.	Author(s)	Title	Key Findings	Our Contribution
1	Breiman, L. (2001) [1]	Random Forests	Introduced random forests, an ensemble learning method for classification tasks.	We leverage the power of random forests for mental health prediction, exploring its effectiveness alongside other algorithms.
2	Cortes, C. & Vapnik, V. (1995) [2]	Support-Vector Networks	Presented Support Vector Machines (SVM) for efficient classification.	We compare SVM performance with Logistic Regression and Random Forest for mental health detection.

3	James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013) [3]	An Introduction to Statistical Learning	Provided a comprehensive overview of statistical learning techniques.	We utilize concepts from statistical learning to guide our feature engineering and model selection for mental health analysis.
4	De Choudhury, M., De Silva, P., Sanjay, S., & Pentland, A. (2013) [4]	Predicting postpartum changes in emotion and behavior via social media	Demonstrated the potential of social media data for predicting mental health changes.	Our study expands on this concept by focusing on mental health disorders in general, not just postpartum changes.
5	De Choudhury, M., Gamst, M., Moreno, M. A., & Morris, M. R. (2014) [5]	Predicting depression via social media	Highlighted the value of social media footprints for early detection of depression.	Our research builds upon this work by exploring a wider range of mental health conditions beyond depression.
6	Gonzalez-Hernandez, G., Success, A. O., & Lopez-Gee, J. M. (2017) [6]	Catching the Patient's Horizon: A Critical Analysis of Advancements in Organic Language Handling of Health-Lined Script	Emphasized the importance of accurate NLP for mental health diagnostics in text data.	We implement NLP techniques like FastText word embedding to process and analyze mental health-related text data effectively.
7	Iroju, O. G., & Olaleke, J. O. (2015) [7]	A Methodical Review on Organic Language Handling in Healthcare	Reviewed NLP applications in healthcare, discussing various algorithms.	Our work delves deeper into specific NLP techniques used for mental health analysis, such as word embedding and sentiment analysis.
8	Singh, J. W., Ma, S., Desai, P., Patel, S., & Choo, J. (2022) [8]	An Infused LSTM Rooted Method for Depressiveness Detection and Analysis	Proposed an LSTM-based method for depression detection.	We compare the performance of our chosen algorithms (Logistic Regression, Random Forest, SVM) with deep learning approaches like LSTMs in future work.
9	Zhang, T., Sun, A., Mao, J., Sun, S., & Liu, Y. (2022) [9]	Natural Language Processing Employed to Mental Ailment Detection: A Storytelling Review	Provided a narrative review of NLP applications in mental health detection.	Our research offers a more focused analysis comparing the effectiveness of various machine learning algorithms for mental health prediction using text data.
10	Nadkarni, P. M., Ohno-Machado, L., & Waegner, C. W. (2011) [10]	Natural Language Processing: An Introduction	Established the foundation for NLP applications in medical informatics.	We build upon this foundation by applying NLP techniques specifically for mental health text analysis and feature extraction.

3. Methodology

The research embarked on a methodical journey encompassing several critical phases to analyze textual data sourced from individuals grappling with anxiety and depression, as well as user comments on Reddit. Initially, data gathering involved meticulous extraction from these diverse sources, ensuring a comprehensive dataset for analysis. Subsequently, the focus shifted towards executing pre-processing techniques to refine the raw text, including converting text to lowercase, eliminating numbers, punctuation marks, and stop words, and tokenizing the text for further analysis. Following pre-processing, the dataset was partitioned into distinct subsets: a training dataset and a test dataset, essential for evaluating model performance accurately.

The subsequent stages involved advanced methodologies aimed at enhancing the interpretability and predictive power of the data. One crucial step entailed incorporating words into the word embedding process, facilitating the transformation of textual data into a numerical representation. Here, the fastest algorithm played a pivotal role, assigning numerical values or vector representations to individual words. This process enabled efficient processing and analysis of textual data, paving the way for subsequent modeling endeavors.

The modeling phase involved configuring the parameters of the analysis model, focusing on optimizing its performance and accuracy. Parameters such as the optimizer and activation function were fine-tuned to effectively capture underlying patterns within the data. Finally, during the assessment stage, the model's performance was rigorously evaluated using metrics such as accuracy, precision, recall, and F1-Score. These metrics provided valuable insights into the model's efficacy in accurately classifying and predicting mental health conditions based on the analyzed textual data. Through this systematic approach, the research aims to contribute to the advancement of mental health assessment methodologies, ultimately facilitating better understanding and management of anxiety and depression.

Our research leverages the power of three machine learning algorithms for mental health disorder prediction: Logistic Regression, Random Forest, and Support Vector Machine (SVM). Each algorithm offers a distinct approach to tackling the classification challenge, providing valuable insights into a user's mental health state.

3.1. Logistic Regression

This workhorse algorithm functions similarly to a statistical balancing scale. It meticulously weighs the influence of various features, such as sadness levels or euphoric episodes, and tilts the scale towards the most probable mental disorder category based on these calculated weights. This approach offers a clear and interpretable decision-making process, allowing researchers to understand which features hold the most weight in the prediction. However, Logistic Regression might struggle with datasets that exhibit intricate and non-linear relationships between features.

3.2. Random Forest

In contrast to the single-handed approach of Logistic Regression, Random Forest thrives on collaboration. It constructs a multitude of decision trees, each resembling a series of yes-or-no questions crafted from the available features. When a new data point, representing a user's assessment, enters the forest, it is navigated through each decision tree. The final prediction hinges on the most frequent outcome across all the individual trees. This ensemble approach empowers Random Forest to handle complex data structures effectively and produce robust predictions. However, the intricate decision-making processes within the forest can make it more challenging to pinpoint the exact reasoning behind a specific prediction.

3.3. Support Vector Machine (SVM)

Imagine a high-dimensional space where each data point, representing a user's assessment, occupies a unique location. SVM excels at creating a clear boundary, known as a hyperplane, within this space. This hyperplane meticulously separates data points belonging to distinct mental disorder categories, ensuring the maximum margin between the categories. During prediction, SVM efficiently determines on which side of the hyperplane the new data point falls, classifying it into the corresponding disorder category. This method shines when the categories are well-defined and possess clear distinctions. However, for datasets where the boundaries between categories might be blurry or less well-established, SVM might not be the most optimal choice.

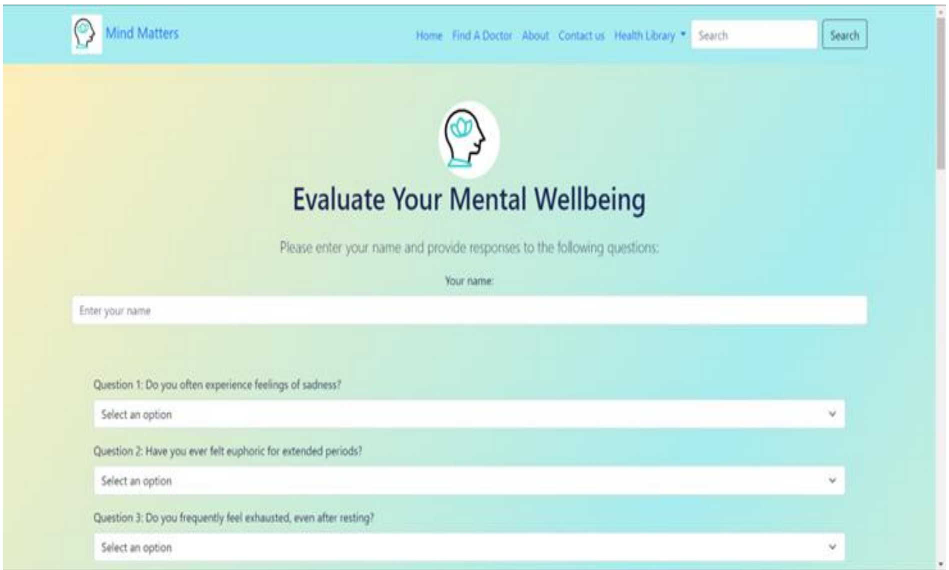

Figure 1. Our website assessment page

Figure 2. Questions on assessment page

4. DATA COLLECTION AND PREPARATION

In the initial phase of our research, we meticulously collected data related to mental health from diverse sources available on Kaggle, a prominent platform for datasets and machine learning projects. Our dataset was sourced primarily from two repositories: the Mental Health Corpus and a Reddit dataset focusing on mental disorders identification, both of which were available on Kaggle. The Mental Health Corpus provided a comprehensive collection of posts with relevant metadata such as timestamps and subreddit information, while the Reddit dataset offered valuable insights into mental health discussions, categorized into distinct labels.

The collected data underwent rigorous preparation to ensure its suitability for our analysis. We organized the dataset into structured formats, removing redundant columns and merging relevant information for comprehensive analysis. Additionally, we partitioned the dataset into distinct subsets, allocating entries for training and testing purposes. This meticulous approach ensured the integrity and reliability of our dataset for subsequent analysis.

Overall, our data collection and preparation processes were meticulously designed to curate a comprehensive dataset suitable for our research objectives. This meticulous approach laid the foundation for our subsequent analysis and model development, facilitating accurate and reliable insights into mental health detection.

5. RESULTS AND DISCUSSION

The performance of various machine learning models was evaluated for the task of detecting Results and Discussion mental health disorders using social media data. The models included Logistic Regression, Random Forest, and Support Vector Machine (SVM), with their Results and Discussion respective accuracy, precision, recall, and f1-scores presented in the following results.

5.1. Logistic Regression

Demonstrated the highest accuracy of 0.92. The precision, recall, and f1-scores for the individual classes (0, 1, 2, 3) were as follows: Class 0 had a precision of 1.00, recall of 0.86, and f1-score of 0.92. Class 1 achieved a precision of 1.00, recall of 0.75, and f1-score of 0.86. Class 2 showed a precision of 0.83, recall of 1.00, and f1-score of 0.91. Class 3 had a precision of 0.89, recall of 1.00, and f1-score of 0.94. The overall macro average for precision, recall, and f1-score were 0.93, 0.90, and 0.91 respectively, and the weighted averages were 0.93, 0.92, and 0.92.

5.2. Random Forest

Achieved an accuracy of 0.79. For Class 0, the precision was 0.80, recall was 0.57, and f1-score was 0.67. Class 1 had a precision of 0.67, recall of 0.50, and f1-score of 0.57. Class 2 obtained a precision of 0.83, recall of 1.00, and f1-score of 0.91. Class 3 had a precision of 0.80, recall of 1.00, and f1-score of 0.89. The macro averages for precision, recall, and f1-score were 0.78, 0.77, and 0.76, while the weighted averages were 0.78, 0.79, and 0.78.

5.3. Support Vector Machine (SVM)

Reached an accuracy of 0.83. The performance metrics for Class 0 included a precision of 1.00, recall of 0.57, and f1-score of 0.73. Class 1 had a precision of 0.75, recall of 0.75, and f1-score of 0.75. Class 2 achieved a precision of 0.83, recall of 1.00, and f1-score of 0.91. Class 3 had a precision of 0.80, recall of 1.00, and f1-score of 0.89. The macro averages were 0.85 for precision, 0.83 for recall, and 0.82 for f1-score, while the weighted averages were 0.86, 0.83, and 0.82 respectively.

In summary, the Logistic Regression model outperformed the other models in terms of overall accuracy and balanced performance across all classes. The Random Forest and SVM models also showed promising results but were less consistent across different metrics. These findings suggest that Logistic Regression may be more effective for this specific task, although further tuning and exploration of model parameters could improve the performance of all models.

Table 1. Logistic Regression Accuracy 0.9166666666666666

Precision	Recall	F1-Score	Support
1.00	0.86	0.92	7
1.00	0.75	0.86	4
0.83	1.00	0.91	5
0.89	1.00	0.94	8

Accuracy			0.92	24
Macro Avg	0.93	0.90	0.91	24

Weighted Avg	0.93	0.92	0.92	24

Table 2. Random Forest Accuracy: 0.7916666666666666

Precision	Recall	F1-Score	Support
0.80	0.57	0.67	7
0.67	0.50	0.57	4
0.83	1.00	0.91	5
0.80	1.00	0.89	8

Accuracy			0.79	24
Macro Avg	0.78	0.77	0.76	24
Weighted Avg	0.78	0.79	0.78	24

Table 3. SVM Accuracy: 0.8333333333333334

Precision	Recall	F1-Score	Support
1.00	0.57	0.73	7
0.75	0.75	0.75	4
0.83	1.00	0.91	5
0.80	1.00	0.89	8

Accuracy			0.83	24
Macro Avg	0.85	0.83	0.82	24
Weighted Avg	0.86	0.83	0.82	24

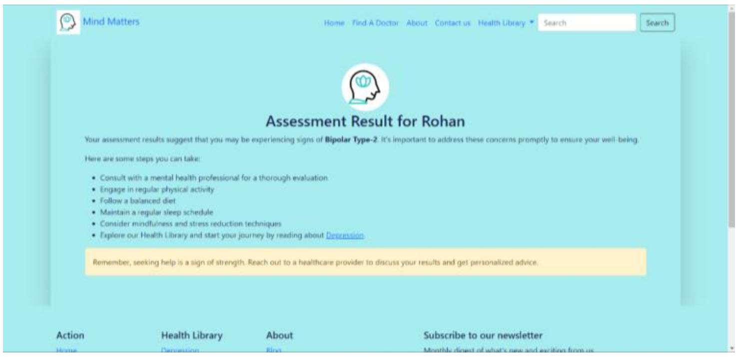

Figure 2(a) assessment result

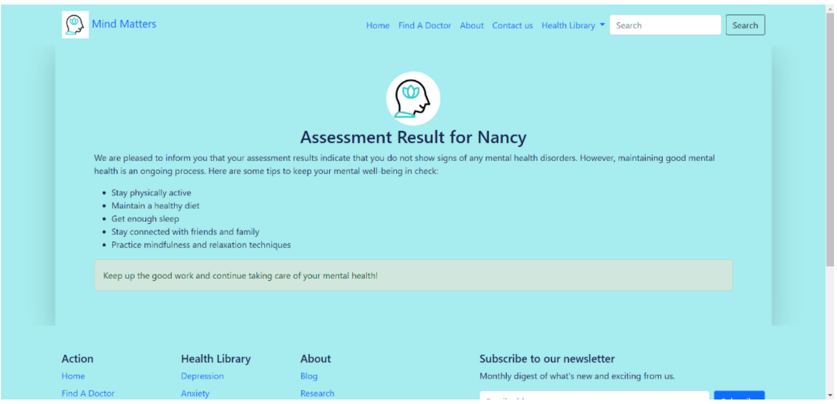

Figure 3(b) assessment result

CONCLUSION

In recent decades, social media has become a prevalent platform for individuals to express their emotional states through various mediums, including text, images, and videos. The COVID-19 pandemic has exacerbated mental health concerns, particularly among those already vulnerable due to isolation measures. There is an urgent need for a detection method to provide timely alerts and aid in diagnosing mental health disorders. To address these challenges, this study proposed the utilization of a model leveraging Fast text's pre-trained word weighting.

The model focused on text data related to mental health, including terms such as borderline personality disorder (BPD), anxiety, depression, bipolar disorder, mental illness, schizophrenia, and self-harm. The evaluation revealed promising results, with the model achieving an accuracy of 86% and an F1-Score of 0.85%. This performance surpasses previous approaches, notably `

However, one notable challenge encountered during the modeling process is the model's difficulty in comprehending the context of lengthy sentences, potentially leading to biased interpretations. Despite this challenge, the model effectively identified patterns associated with mental health, highlighting its potential to enhance detection and diagnosis in this critical area.

REFERENCES :

[1] L. Breiman, "Random forests," *Machine Learning,* pp. 5-32, 2001.

[2] C. &. V. V. Cortes, "Support-vector networks," *Machine Learning,* pp. 273-297, 1995.

[3] G. W. D. H. T. &. T. R. James, An Introduction to Statistical Learning: with Applications in R. Springer Science & Business Media, 2013, p. 662.

[4] M. D. S. P. S. S. &. P. A. De Choudhury, "Predicting postpartum changes in emotion and behavior via social media. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining.," 2013.

[5] M. G. M. M. M. A. &. M. M. R. De Choudhury, "Predicting depression via social media. First Monday, 19(8).," 2014.

[6] G. S. A. O. &. L.-G. J. M. Gonzalez-Hernandez, "Catching the Patient's Horizon: a Critical Analysis of Advancements in Organic Language Handling of Health-Lined Script," *Yearbook of Medicinal Details,* pp. 235-239, 2017.

[7] O. G. &. O. J. O. Iroju, "A Methodical Review on Organic Language Handling in Healthcare," *Intercontinental Journal of Informative Tech and Computer Studies,* pp. 102-108, 2015.

[8] J. W. M. S. D. P. P. S. &. C. J. Singh, "An infused LSTM rooted method for depressiveness detection and analysis," *Sensors,* pp. 166-175, 2022.

[9] T. S. A. M. J. S. S. &. L. Y. Zhang, "Natural Language Processing Employed to Mental Ailment Detection: A Storytelling Review," *Brain sciences,* pp. 1-13, 2022.

[10] P. M. O.-M. L. &. W. C. W. Nadkarni, Natural Language Processing: An Introduction, John Wiley & Sons, 2011, pp. 544-551.

[11] M. C. S. &. H. E. De Choudhury, "Predicting postpartum changes in emotion and behavior via social media," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* pp. 3267-3276, 2013.