Depression Severity Classification from Social Media Text using Natural Language Processing and Machine Learning

Abhijeetsinh Jadeja
Sankalchand Patel University, Visnagar, India

Priyanka Ameta Geetanjali, PhD
Institute of Technical Studies

Deepika Ameta
IIMHRD, Pune

Asha Patil
R. C. Patel Institute of Management Research and Development Shirpur, Dhule (MH)

ABSTRACT
With its potential for early diagnosis, research on mental health monitoring is an active area and automatic analysis is an important component of such a system. However most research involves simply detecting presence/absence of depression, which is not sufficiently granular for practical application. We propose the development of a interactive chatbot which would classify user responses into four severity levels of depression- Minimal, Mild, Moderate and Severe. We developed an NLP pipeline using lemmatization and TF-IDF vectorization to train and compare a Logistic Regression model with a fine-tuned Support Vector Machine. Results indicate that the SVM model achieved 74.36% accuracy among other algorithms and could be used as a suitable engine to provide an interactive conversational interface to asses user's current stress level in real time.

Keywords
Depression Detection, NLP, Machine Learning, Chatbot, Severity Classification, TF-IDF, SVM, SMOTE.

1. INTRODUCTION
Though the estimated 350 million individuals worldwide are suffering from depression, almost 70% individuals in early stage of disorder never receive professional care due to the long-lasting stigma related to mental disorder and the lack of readily available diagnosis resources. This gap between the need of medical attention and its access cause individuals to bear huge loss of personal life and economy across the world. Early identification of symptoms are crucial to not only policy matters but is essential for life saving purpose where one would direct risk person to the suitable prevention tool before an accident happens.

Although, modern diagnosis method involves clinical interview and static documents, modern NLP allows for more flexible "human voice" understanding by analyzing the way one usually expresses his thoughts and feelings. Such models often achieve greater accuracy than any existing predictive models. Unfortunately, many current literature approach depression in a "yes/no" binary fashion which fails to distinguish severity level and also fails to provide helpful information for the clinical triaging.

In reality, mental disorder falls in a continuum where each level requires its respective intervention for diagnosis and care. By allowing users to be interactive and responsive to a conversation with the chatbot and then performing analysis of the conversation with the machine learning model, users will find themselves assigned a severity level for their particular stress level rather than simply have their social media passively mined for an indication of depression. This combination provides a dynamic system for detecting severity level through interaction which modern age of modern world require for a smarter medical assistance system.

2. RELATED WORK
Depression is a highly prevalent mental health condition, affecting a significant portion of the world population. With the advent of machine learning and artificial intelligence, substantial work has been dedicated to the automated detection and estimation of the severity of depression using clinical, behavioural, and digital data.

Both structured and unstructured data sources in the prediction of depression have also been used in supervised machine learning. Conventional algorithms like Logistic Regression, SVMs, K-Nearest Neighbours (KNNs), and Random Forests have been applied extensively for depression classification tasks [3][5][13], and in one study, Word2Vec embeddings of social media text and Random Forest have delivered an approximate 0.877 accuracy and F1-score [1].

Survey papers have been used to give a good overview of the diverse datasets that are available, their pre-processing procedures, and the feature extraction methods used in depression studies from social media and digital platforms [2][19]. There are also speech-based depression detection systems that have been developed using acoustic characteristics such as MFCC and pitch to forecast severity with excellent success [3].

These methods play a significant role in early diagnosis and follow up. Patient Health Questionnaire (PHQ-9) is an established clinical measure on depression severity. This is a nine-item questionnaire offering an ordinal scale of 0 to 27, which is the ground truth in training machine learning models to severity into discrete classes (mild, moderate, severe). Formatted data, such as the PHQ-9 Student Depression Dataset, have been extensively used to train models, such as the Random Forest and Support Vector Machines (SVMs) to obtain depression severity classification [4].

Over the last several years numerous studies employ AI and machine learning to identify depression or levels of depression. This assists in the diagnosis of the disorder and can be used to monitor depression within large scale population. PHQ-9 scale is one of the most popular clinical standards that measure the level of depression as mild to severe. Prediction was done using machine learning models like Logistic Regression, SVM, Random Forest and KNN. Also, other studies used multimodal sources of data such as smartphone sensors and user behaviour.[5]. In one study, ordinal classification was used to

determine the level of depression severity on Reddit and combine attention mechanisms and ordinal loss functions to predict the severity of the disorder at different levels with a higher level of accuracy [6].

Moreover, it has been noted that the fine-grained and multi-label classification of symptoms have gained some interest. The DepressionEmo dataset enables emotion multi-labeling of feelings in a Reddit comment via BART and XGBoost models, thus describing co-occurring affective states [7]. In addition, the HelaDepDet data set and corresponding models of the multi-class depression severity classification were trained on more than 40,000 annotated social media statements, which performed better at the many levels of the severity of depression [8].

Equally, the RESTORE dataset incorporates multimodal information of the memes (text and images) to identify fine-grained symptoms in the PHQ-9 categories, which underscores the significance of the visual and textual information combination [9].

Natural Language Processing (NLP) has played a significant role in the analysis of social media posts on social media platforms such as Twitter and Reddit. Systems such as DEPTWEET were able to offer a dataset with various degrees of depression severity and used transformer-based models such as BERT to classify tweets [10]. Knowledge from broader NLP applications, including named entity recognition and text classification studies, has further contributed to advancements in mental health analytics and social media text understanding [14][15][16][17][18].

The multimodal techniques have expanded the assessment boundaries of depression. Deep-learning based and sentiment analysis of Flickr posts during the COVID-19 pandemic showed how the emotional expression and behaviours changed depending on a combination of text and image data [11]. Moreover, deep learning algorithms such as GRU and LSTM have demonstrated better results than the traditional algorithms in identifying severity of posts on Bengali social media with a reported accuracy of approximately 81% [12].

Recent studies have further contributed to the advancement of machine learning applications in text analytics and depression detection. N-gram based feature extraction combined with ensemble learning has demonstrated effectiveness in classification tasks [20]. Sentiment analysis and text mining techniques have been widely used to extract emotional patterns from textual data [22], while data preprocessing strategies, including the handling of missing values, have been shown to improve model reliability and performance [23]. Multimodal approaches that combine visual and textual information have also enhanced depression recognition accuracy [24]. These developments highlight the growing importance of robust feature engineering, data quality management, and multimodal learning in automated mental health assessment systems [20]–[24]. Advances in NLP have enabled the creation of realistic and context-aware textual content for machine learning applications, demonstrating the versatility of NLP techniques beyond traditional text classification tasks [25].

3. METHODOLOGY

We follow a pipeline from raw data processing all the way down to a functional classification model. This workflow shows how raw user input is transformed by first transforming it into a preliminary stage of work, such as text normalization, lemmatization, and TF-IDF vectorization, towards the models that finally made the classification. Linear SVM model was chosen as the final inference engine to maximize the performance because it is more accurate than the Logistic Regression.

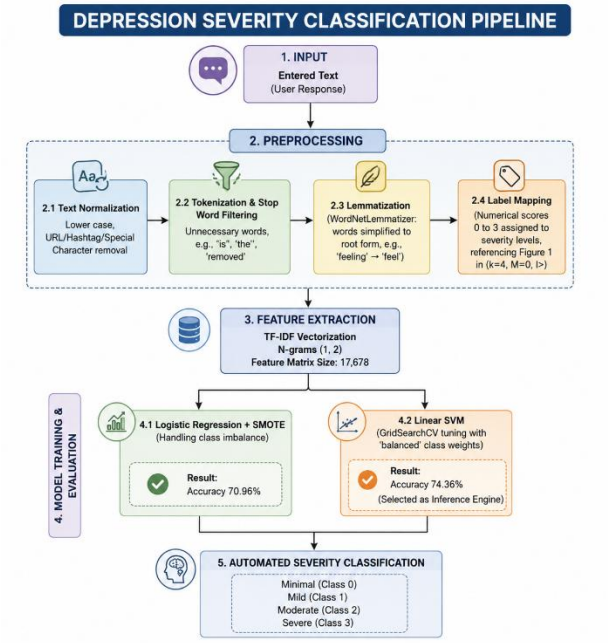

Figure 1: Depression severity classification using automated text analysis proposed NLP pipeline

3.1 Data acquisition and cleaning

The dataset used in this study was obtained from a publicly available GitHub repository containing mental health related social media posts. The dataset comprises approximately 3,500 Reddit-style textual posts generated by users discussing their emotional and psychological experiences. Each post is annotated with one of four depression severity levels such as Minimal, Mild, Moderate, and Severe. The dataset characteristics are shown in Table I.

Table I: Dataset Characteristics

Dataset Attributes	Description
Source	Public GitHub Repository
Samples	3500 Posts
Data Type	Textual (Reddit Style Posts)
Labels	Minimal, Mild, Moderate, Severe

The model has been trained with an organized set of queries that are given multi class labels. Subsequent data cleaning procedures were done to prepare the data that was to be used by the machine learning model.

Text Normalization- The entire input was normalized to lower case and all the noise words such as URLs, hashtags, special characters were eliminated.

Tokenization and stop word filtering- All the unnecessary words such as is, the etc were eliminated to ensure that model focuses on words that have high sentiment.

Lemmatization- Words were simplified to their root word with WordNetLemmatizer such that we have an identical word such as feel to feeling (varied tenses).

Label Mapping- The rankings of the severity categorization were converted to a computational form with each level being allocated a score between 0 and 3.

3.2 Feature extraction

The text input vectorised using TF-IDF to convert them to numerical format that computer can process. The choice of n-gram was (1, 2) where 1 refers to words, 2 refers to words and pair of words. With this feature matrix, we get a feature set size of 17,678, enabling us to capture all subtle linguistic patterns relevant to different emotions.

3.3 Model Performance

The experiment with two machine learning methods for classification is implemented and analyse them. The model performance for our is shown in Table II.

1. **Logistic Regression with SMOTE**- Since mental health data generally involves class imbalance, SMOTE (Synthetic minority oversampling technique) was used to synthetically over sample the minority classes (Mild, Moderate and Severe). It yields an accuracy of 70.96%.
2. **Linear SVM**- Support Vector Machine was tuned using GridSearchCV with class weights set to 'balanced' to penalize minority classes correctly. It yielded a better accuracy of 74.36%, and we select it as the final inference engine.

Table II: Model Performance

Computational Model	Class Imbalance Handling	Accuracy (%)
Logistic Regression with SMOTE	SMOTE	70.96
Linear SVM	Class Weight = Balanced + GridSearchCV	74.36%

3.4 Result Analysis

The SVM was able to effectively identify those users with "Minimal" depression and has a high recall of 0.94. Although the model had difficulty discriminating between closely related classes (e.g. Mild/Moderate), the Macro F1-score of 0.433 suggests that the model is a solid, reliable starting point for an automated screening tool.

The model excels at identifying those with "Minimal" depression (Class 0) and has a high recall of 0.94. This suggests that this system can serve effectively as a "negative" screen as it rarely misclassifies users with no or mild depression symptoms.

The precision and recall for mild (Class 1), moderate (Class 2), and severe (Class 3) categories are significantly lower (0.19-0.43). The lower scores are due to a number of factors: Overlapping language categories. The words used to describe mild, as opposed to moderate, depression are, at times, semantically identical. It is very difficult for a model based on TF-IDF to draw a precise line. Imbalanced data, Despite balancing the data with class weighting and SMOTE in training, the predominance of users in "Minimal" category seems to take priority.

This is because looking at the confusion matrix of the training, we notice that there are more false classifications between closely related classes. False Positives. The moderate but not mild cases were always categorized as minimal depression. The line is thin. Class 1 and class 2 appear to be conditional upon an intensity of a symptom (sometimes, most times, all times). In lemmatization this is often lost.

5. FUTURE WORK

The current paper provides a strong platform on which automatic severity classification can be achieved, yet still it has a lot of room to enhance the clinical validity of the model, and the reactivity of the user. Multimodal data inclusion: Future versions may advance beyond text only data including acoustic data (pitch, tone, pauses), the facial recognition through the camera to provide more of the general diagnostic picture.

Temporal and Sequential analysis: The model would be created so that it would monitor the mental condition of a patient over time rather than a single chat. Analyzing previous data and considering the progression of events in which mental states are likely to shift may possibly reveal incidences that may result in high risk scenarios.

State-of-the-art Transformer models: In addition to common ML models, like SVMs, it is possible to switch to modern transformer models, like BERT or RoBERTa, which may be able to raise the F1-score of the minority classes of severity by being able to capture more context effectively in a human speech, as observed in the Bengali study where the transformer demonstrated better scores.Greater Culture and Language Diversity: It is currently being streamlined to work well with the English language and it can be invaluable to consider regional languages such as in the Bengali study. This would expand the model to be used across cultures and places. Anticipatory crisis intervention: Future emphasis will also be placed on developing a Red Flag system in the present moment; in case of a chatbot, highly rating the reply as being of the type of Severe, it would automatically leave links to hotlines or contact sources.

Clinical assessment and incorporation: Consultation with mental health care providers about the most appropriate means to implement this in patient portals would streamline the triage process of mental health care to enable screening and professional psychiatric care.

6. CONCLUSION

The adoption of the "Depression Severity Analysis Chatbot" is a step in the right direction in terms of accessibility and detailedness of mental health screening tools. This tool can provide a better degree of specificity by abandoning binary classification and transitioning to four levels of severity (Minimal, Mild, Moderate, Severe) which is more consistent with established clinical guidelines such as PHQ-9 and BDI-II.

Our tests have shown that a hybrid design, using the standard Natural Language Processing and effective classifiers of Machine Learning, can be used to give a reliable classification system. The Support Vector Machine (SVM) model that had a weighted class feature and TF-IDF vectorizer gave the best system backend and served as the main classification engine with an overall accuracy of 74.36%. Although the model was especially effective in the prediction of the cases that belonged to the Minimal category (recall 94 percent), the fact that the distinct categories cannot be further differentiated (e.g., Mild and Moderate) can be attributed to the heavily convoluted nature of the human emotional expression.

The outcome of the project is a working interactive chatbot that has converted our backend logic into a user experience. The chatbot does not give a single result, which results from a fixed analysis of the data, but gives the user a live evaluation of the current stress levels. The Depression Severity Analysis Chatbot is not an alternative to medical diagnosis, but a first point of contact, where it relies on a data-driven strategy with an emotional overture. Its capability to rapidly expand and offer a further degree of privacy will most probably make it a helpful element in the endeavours to reduce the global burden of undiagnosed depression.

7. REFERENCES

[1] Sampath, K., and Durairaj, T. 2022. Data Set Creation and Empirical Analysis for Detecting Signs of Depression from Social Media Postings. *International Conference on Computational Intelligence in Data Science*, pp. 136–151. Springer International Publishing.

[2] Bucur, A. M., Moldovan, A., Parvatikar, K., Zampieri, M., Khudabukhsh, A., and Dinu, L. P. 2025. Datasets for Depression Modeling in Social Media: An Overview. *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pp. 116–126.

[3] Hassan, M. M., Khan, M. A. R., Islam, K. K., Hassan, M. M., and Rabbi, M. F. 2021. Depression Detection System with Statistical Analysis and Data Mining Approaches. *International Conference on Science and Contemporary Technologies (ICSCT)*, pp. 1–6. IEEE.

[4] Li, J., Luo, C., Liu, L., Huang, A., Ma, Z., Chen, Y., and Zhao, J. 2024. Depression, Anxiety, and Insomnia Symptoms Among Chinese College Students: A Network Analysis Across Pandemic Stages. *Journal of Affective Disorders*, 356, pp. 54–63.

[5] Hussain, N., Qasim, A., Mehak, G., Zain, M., Sidorov, G., Gelbukh, A., and Kolesnikova, O. 2025. Multi-Level Depression Severity Detection with Deep Transformers and Enhanced Machine Learning Techniques. *AI*, 6(7), 157.

[6] Naseem, U., Dunn, A. G., Kim, J., and Khushi, M. 2022. Early Identification of Depression Severity Levels on Reddit Using Ordinal Classification. *Proceedings of the ACM Web Conference 2022*, pp. 2563–2572.

[7] Priyadarshana, Y. H. P. P., Liang, Z., and Piumarta, I. 2023. HelaDepDet: A Novel Multi-Class Classification Model for Detecting the Severity of Human Depression. *International Conference on Collaboration Technologies and Social Computing*, pp. 3–18. Springer Nature Switzerland.

[8] Kabir, M., Ahmed, T., Hasan, M. B., Laskar, M. T. R., Joarder, T. K., Mahmud, H., and Hasan, K. 2023. DEPTWEET: A Typology for Social Media Texts to Detect Depression Severities. *Computers in Human Behavior*, 139, 107503.

[9] Fernández-Barrera, I., Bravo-Bustos, S., and Vidal, M. 2022. Evaluating the Social Media Users' Mental Health Status During COVID-19 Pandemic Using Deep Learning. *International Conference on Biomedical and Health Informatics*, pp. 60–68. Springer Nature Switzerland.

[10] Kabir, M. K., Islam, M., Kabir, A. N. B., Haque, A., and Rhaman, M. K. 2022. Detection of Depression Severity Using Bengali Social Media Posts on Mental Health: Study Using Natural Language Processing Techniques. *JMIR Formative Research*, 6(9), e36118.

[11] Fernández-Barrera, I., Bravo-Bustos, S., and Vidal, M. 2022. Evaluating the Social Media Users' Mental Health Status During COVID-19 Pandemic Using Deep Learning. *International Conference on Biomedical and Health Informatics*, pp. 60–68. Springer Nature Switzerland.

[12] Kabir, M. K., Islam, M., Kabir, A. N. B., Haque, A., and Rhaman, K. 2022. Detection of Depression Severity Using Bengali Social Media Posts on Mental Health: Study Using Natural Language Processing Techniques. *JMIR Formative Research*, 6(9), e36118.

[13] Patil, J., Patil, V., Prajapati, K., Patel, D., Trivedi, S., and Patel, R. 2024. Enhanced Depression Detection on Social Media Using Advanced Machine Learning and Linguistic Analysis Techniques. *International Conference on Intelligent Computing and Communication*, pp. 263–275. Springer Nature Singapore.

[14] Verma, S., Patil, J. A., and Tamhankar, I. 2024. Integrating Natural Language Processing with CGANs to Generate Customized, Realistic Traffic Scenarios for Autonomous Vehicle Training. *IEEE International Conference on Smart Power Control and Renewable Energy (ICSPCRE)*, pp. 1–5.

[15] Patil, M. J. A., and Godhwani, M. P. B. 2016. Review of Name Entity Recognition in Marathi Language. *IJSART*, 2(6).

[16] Patil, J., and Sheth, J. 2022. Deep Learning and Machine Learning Approaches for the Classification of Personality Traits. In *Rising Threats in Expert Applications and Solutions: Proceedings of FICR-TEAS 2022*, pp. 139–146. Springer Nature Singapore.

[17] Patil, J., and Sheth, J. 2021. Comparative Study of Data Sources, Features, and Approaches for Automatic Personality Classification from Text. *International Journal of Computer Applications*, 174.

[18] Patil, J., and Sheth, J. 2022. Data Preparation and Quality Challenges for Personality Recognition in Indian Languages Using Machine Learning and Deep Learning Approaches. *Journal of IoT in Social, Mobile, Analytics, and Cloud*, 4(1), pp. 33–40.

[19] Patil, J., et al. 2025. A Review of Transforming AI for Depression Detection: Transformer Model Dominance, Multimodal Approaches, and Future Pathways. *International Conference on Computing and Machine Learning*. Springer Nature Singapore.

[20] Pandya, D. D., Jadeja, A., Degadwala, S., and Vyas, D. 2022. Ensemble Learning Based Enzyme Family Classification Using N-Gram Feature. *6th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 1386–1392. IEEE.

[21] Pandya, D. D., Jadeja, A., Degadwala, S., and Vyas, D. 2023. Retraction Notice: Diagnostic Criteria for Depression Based on Both Static and Dynamic Visual Features. *International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, p. 1. IEEE.

[22] Pandya, D., Jadeja, A., Khan, M. A., Trivedi, S. B.,

Ramnath, M. A., and Satish, B. P. 2024. Significance of Sentiment Analysis with Text-Based Mining Approach. *International Conference on Emerging Trends in Expert Applications and Security*, pp. 315–323. Springer Nature Singapore.

[23] Pandya, D., Jadeja, A., Gour, S., Trivedi, S. B., Patel, H. H., and Jadeja, P. U. 2024. An Analytical Perspective of Missing Values in Machine Learning. *International Conference on Emerging Trends in Expert Applications and Security*, pp. 285–294. Springer Nature Singapore.

[24] Kacha, H., Bhikadiya, D., Sharma, K., and Patil, J. 2025. Comparative Analysis of Visual Motion and Multimodal Strategies in Depression Recognition. *International Journal of Computer Applications*, 187(58), pp. 58–64.

[25] Verma, S., Patil, J. A., & Tamhankar, I. (2024, July). Integrating Natural Language Processing with CGANs to Generate Customized, Realistic Traffic Scenarios for Autonomous Vehicle Training. In *2024 IEEE International Conference on Smart Power Control and Renewable Energy (ICSPCRE)* (pp. 1-5). IEEE.