



Explainable Machine Learning for Mental Health Detection Using NLP

Noor Ul Ain Mushtaq^{a,b}, Sanam Narejo^a, Syed Amjad Ali^b,
and Muhammad Moazzam Jawaid^c

^a Department of Computer System Engineering, Mehran University of Engineering and Technology, Indus Hwy, Jamshoro, Sindh 76062

^b Department of Biomedical Engineering, Mehran University of Engineering and Technology, Indus Hwy, Jamshoro, Sindh 76062

^c School of Engineering & Physical Sciences, University of Lincoln, Lincoln LN6 7TS, United Kingdom

✉ noorulain.mushtaq@admin.muet.edu.pk, sanam.narejo@faculty.muet.edu.pk,
amjad.ali@faculty.muet.edu.pk, MJawaid@lincoln.ac.uk

KEYWORDS

mental health;
machine learning
(ML); black-box
models; explainable
AI (XAI); natural
language processing
(NLP); LIME (local
interpretable model-
agnostic explanations)

ABSTRACT

Humans' mental conditions are often revealed through their social media activity, facilitated by the anonymity of the internet. Early detection of psychiatric issues through these activities can lead to timely interventions, potentially preventing severe mental health disorders such as depression and anxiety. However, the complexity of state-of-the-art machine learning (ML) models has led to challenges in interpretability, often resulting in these models being viewed as «black boxes». This paper provides a comprehensive analysis of explainable AI (XAI) within the framework of Natural Language Processing (NLP) and ML. Thus, NLP techniques improve the performance of learning-based methods by incorporating the semantic and syntactic features of the text. The application of ML in healthcare is gaining traction, particularly in extracting novel scientific insights from observational or simulated data. Domain knowledge is crucial for achieving scientific consistency and explainability. In our study, we implemented Naïve Bayes and Random Forest algorithms, achieving accuracies of 92 % and 99 %, respectively. To further explore transparency, interpretability, and explainability, we applied explainable ML techniques, with LIME emerging as a popular tool. Our findings underscore the importance of integrating XAI methods to better understand and interpret the decisions made by complex ML models.

1. Introduction

In recent years, the analysis of stress and sentiment in posts on microblogging platforms has advanced significantly. Given that mental stress can lead to numerous health issues, early detection and intervention are crucial. High levels of mental stress often manifest in both mental and physical behavioral changes (American Psychological Association, 2021). Traditionally, stress analysis has utilized data from platforms such as Twitter and Facebook. This study aims to explore global stress factors and characteristics by performing stress analysis and identifying stress indicators.

The objective is to use these insights to enhance the conversation around mental health and support early intervention for conditions such as depression. There is an evident need for systems that can detect stress and mental disorders early. Social media platforms serve as crucial outlets for emotional expression, and the extensive text posts available can help identify signs of stress and mental distress.

The importance of explainability in AI has become increasingly recognized, giving rise to the field of explainable AI (XAI). Just as different academic disciplines require various methodologies, the study of explainability varies across different contexts. The widespread adoption of machine learning (ML) in commercial applications, especially neural networks (NNs), has led to its significant use in scientific research as shown in Figure 1. Although these models are often fine-tuned for high accuracy, there is a growing need to understand their operations and the reasons behind their decisions. This understanding is critical as ML becomes more integral to scientific optimization and discovery. Consequently, explainability is vital for validating scientific outcomes. Research areas such as explainable artificial intelligence (XAI) (Samek & Müller, 2019), informed ML (Von Rueden et al., 2021), and intelligible intelligence (Weld & Bansal, 2019) have emerged, each with unique concepts, goals, and definitions. In this paper, we apply machine learning algorithms, specifically

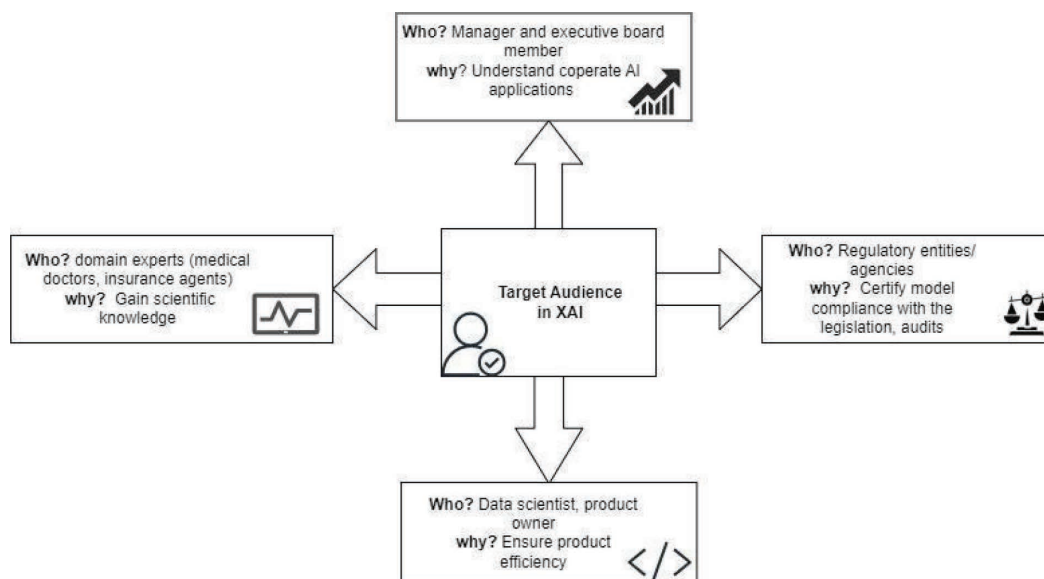


Figure 1. Diagram representing the different purposes of explainability in ML models sought by different audience profiles

Naive Bayes (NB) and Random Forest (RF), to identify stress indicators in social media posts. Our study includes text preprocessing and exploratory data analysis (EDA) as part of the NLP component to prepare and analyze the data. The key contribution of our research is the notable results obtained by combining these algorithms with explainable machine learning techniques for text classification and stress detection. We also examine one of the most widely used interpretability methods: Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016), focusing on its application to text data. For example, LIMEs algorithm to explain the prediction of model f , can be summarized as follows:

- (i) From a corpus of documents C , create a TF-IDF transformer ϕ embedding documents into RD
- (ii) Create n perturbed documents $x_1, \dots, x_n$ by deleting words at random in ξ
- (iii) For each new example, get the prediction of the model $y_i = f(\phi(x_i))$;
- (iv) Train a (weighted) linear surrogate model with inputs, the absence or presence of words and responses denoted by y_i s.

The widespread popularity of LIME can be attributed to its model-agnostic approach, which allows users to generate explanations without needing to understand the specific architecture of the model f . It is sufficient to be able to make numerous queries to the model. Additionally, the explanations provided are highly intuitive, enabling users to easily verify whether the model is performing correctly (or not) on a given example.

This research begins by reviewing earlier studies on stress analysis, while Section 2 delves into the outcomes that the authors had previously attained. Subsequently, Section 3 presents the proposed method of experimentation on the dataset and the details about the dataset. This section also describes the models used and the approach to this problem. Furthermore, Section 4 demonstrates the results obtained by the authors. In Section 5, the authors present the conclusion, future scope, and limitation.

2. Literature Review

2.1. Analyzing Depression Using Social Media Data

This section offers a comprehensive overview of the research on the analysis of depression using language processing methods on social media data. Over the past decade, the field of mental health diagnosis using social media has expanded significantly. Analyzing word usage in a variety of situations, including poetry (Stirman & Pennebaker, 2001), college essays (Rude et al., 2004), and participant narrative styles (Zinken et al., 2010), has been used in previous studies to investigate mental diseases. For example, a paper from Tsinghua University in China considered multiple behavioral factors when constructing a dataset from an existing collection of Twitter posts, focusing on developing a hybrid detection model.

In recent years, social media has increasingly become a platform where individuals express themselves. Research utilizing social media data to identify toxic speech with Deep Learning methods, as well as detecting self-reported signs of depression through neural networks (Kim et al., 2020), has achieved significant advancements in natural language processing (NLP). These studies categorize users' mental health issues according to the posts they make in particular categories or subreddits.

Another study by Pirina et al. (2018) focused on identifying depression by examining user activity and interactions. Other research has addressed the connection between chronic stress-related illnesses

and urban settings (Sekulić & Strube, 2020), and they have found that a large number of patients with these illnesses share their experiences on social media, either directly or indirectly. Additionally, the mental state known as Fear of Missing Out (FOMO) in adolescents has been linked to emotional symptoms that lead to chronic or acute stress-related conditions (Fabris et al., 2020). The same study highlighted that FOMO is closely related to users' sensitivity to social media addiction, which increases their need to consume more content in search of inspiration or self-confidence. However, this often results in diminished self-esteem and further deterioration of mental health (Brailovskaia et al., 2020).

There are also challenges associated with accessing social media datasets, as they often require anonymization. Platforms such as Reddit, whose users are typically anonymous, provide a handier data source even though this process might be laborious. When examining social media posts to find patterns suggestive of stress or other mental health disorders, Guntuku et al. (2017) took into account a number of criteria. Their dataset, which included only 600 Americans, was less illustrative of larger worldwide patterns. Psychological habits can be greatly influenced by cultural differences, so what works in one nation might not work in another.

In a separate study, Lin et al. (2014) proposed a deep learning model to detect psychological stress in users. In addition to the postings' reach or scope, they concentrated on content elements such as text and photos. The challenge of merging these disparate properties was addressed by the authors using autoencoders. While their feature extraction method from posts is innovative, the authors did not use a labeled dataset, opting instead to label their data using linguistic attributes, which can sometimes miss underlying characteristics.

2.2. Machine Learning Models

In a recent study, Myles et al. (2004) investigated various well-known machine learning algorithms, including Decision Trees (DT) and Random Forests (RF), which utilize tree traversal methods for making predictions or classifications. In essence, these models operate like a series of if-else statements. This implies that they can only assign a label to a data point if all relevant features are included in the training dataset.

Another study, by Yong et al. (2009), addressed challenges with the k-nearest neighbor (KNN) algorithm when dealing with text data. The issue stems from the many different ways to describe the same situation in English, which complicates the preprocessing stage and prevents the elimination of perspective-related problems. As a result, plotting data points on an n-dimensional graph for KNN analysis is often ineffective. Furthermore, the instance-based nature of KNN makes it impractical for larger datasets (Guo et al., 2003). Nevertheless, certain techniques can help reduce vector size and enhance accuracy (Gayathri & Marimuthu, 2013).

The Support Vector Machine (SVM) is known for its high performance, but one of its drawbacks is lower recall, which can be a limitation in certain applications (Lodhi et al., 2002). Despite this, SVM classifiers tend to deliver superior results in many classification tasks, particularly in disease identification and face recognition (Abdullah & Abdulazeez, 2021). In the context of predicting chronic kidney disease, three machine learning algorithms, Random Forest, k-Nearest Neighbors (kNN), and Naïve Bayes were evaluated, with Random Forest emerging as the most effective (Sinha & Sinha, 2015; Ibrahim & Abdulazeez, 2021). This success can be attributed to the use of random sampling and ensemble strategies, which lead to more accurate predictions and better generalization (Matta & Saraf, 2020).

Machine learning techniques are increasingly crucial in the medical field, not only for analyzing diseases from medical records but also for addressing challenges such as those posed by the COVID-19 pandemic. These techniques have been instrumental in understanding public perceptions and misconceptions

during the pandemic (Becker et al., 2022). Moreover, machine learning is becoming an indispensable tool in drug discovery, as it enables the analysis of existing drug data to predict new needs with its usage growing rapidly (Elbadawi et al., 2021). For healthcare organizations and professionals, leveraging machine learning to gain insights into what causes anxiety, stress, and trauma among the public is vital for developing better policies and treatments (Praveen et al., 2021). Table 1 summarizes the various machine learning models and their findings.

Table 1. Systematic Literature Review of Machine Learning Models

Title	Author	Methodologies	Findings
Efficient English Text Classification Using Selected Machine Learning Techniques [29]	Luo, X. (2021)	SVM, Naïve Bayes, Logistic Regression	Precision, Recall, and F1-value are calculated for evaluation; SVM outperforms the others in two datasets, and Logistic Regression outperforms in one dataset.
Comparative Analysis of Machine Learning Algorithms on Different Datasets [28]	Uysal & Öztürk (2018)	Neural Network, K-Nearest Neighbor, Support Vector Machine	SVM outperforms the other algorithms; the model is useful in medication, governmental issues, and different fields.
Text Classification Using Machine Learning Techniques [30]	Ikonomakis, Kotsiantis & Tampakas (2005)	Naïve Bayes, K-Nearest Neighbour, Support Vector Machine	Classification performance depends on training text corpora; higher-quality training improves performance.

2.3. Explainable Artificial Intelligence

In recent years, there has been significant growth in the development and deployment of explainable AI (XAI), as noted by Joshi et al. (2021), deep neural networks have become essential in computer vision, natural language processing, and various other fields. Researchers are now exploring the integration of Multimodal AI with XAI to enhance model interpretability and understanding. Al Hammadi et al. (2021) presented a framework that integrates explainability with deep learning and machine learning models using EEG signals to identify industrial insider threats. Hu et al. (2022) proposed a machine learning model for predicting the readmission of septic patients in the Intensive Care Unit, employing SHAP values to extract key features for precise predictions and LIME to provide explanations of the model's outputs. Slijepcevic et al. (2021) developed a method called Layer-wise Relevance Propagation combined with XAI to analyze clinical gait patterns in time series data.

3. Data and Methodology

This section describes the dataset used, and the detailed methodology of the proposed research idea.

3.1. Data Description

There are two columns in this dataset: ‘Text’ and ‘Label.’ The ‘Text’ column contains normal and anxiety/depression-related content, while the ‘Label’ column indicates whether the corresponding text represents anxiety or depression. Table 2 presents the distribution of text and corresponding labels.

Table 2. Example of Depressed and Not-Depressed post

Data item	Text	Label
0	Oh my gosh	1
1	trouble sleeping, confused mind, restless, heart. All out of tune	1
2	All wrong, back off dear, forward doubt. Stay in a restless and restless place	1
.		
6281	tired of clowns but still hopefully tonight if not tomorrow TM but mas tm also no teaser yet....	0

The dataset initially was unbalanced and data augmentation techniques were employed to achieve a balanced distribution.

3.2. Methodology

The methodology is structured in three key steps: natural language processing (NLP), where we preprocess and analyze the text data; Machine Learning (ML), where we build and train models to classify the data; and explainable AI (XAI), where we ensure that the model’s decisions are interpretable and transparent, providing insights into the factors influencing the predictions. The research pipeline is shown in Figure 2.

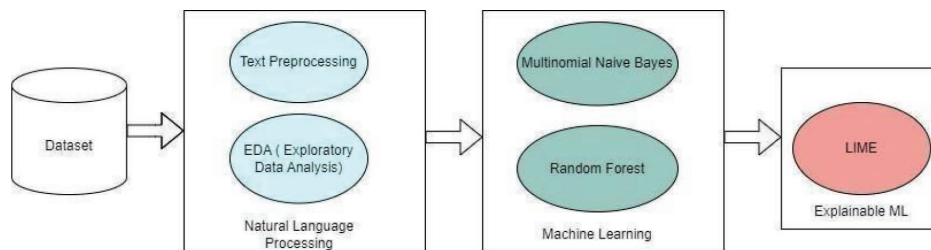


Figure 2. Research pipeline

3.2.1. Natural Language Processing

3.2.1.1. Data Preprocessing

Initially, in the preprocessing stage as shown in Figure 3, the first step performed was tokenization, breaking down the text into individual words or tokens. In the next step, all characters were converted

to lowercase for standardization, followed by removing any URLs to eliminate irrelevant noise. Moreover, punctuations were removed to focus on the meaningful content and common stop words that do not contribute significant meaning were filtered out. Finally, lemmatization was applied to reduce words to their base or root form, ensuring consistency across different word forms.

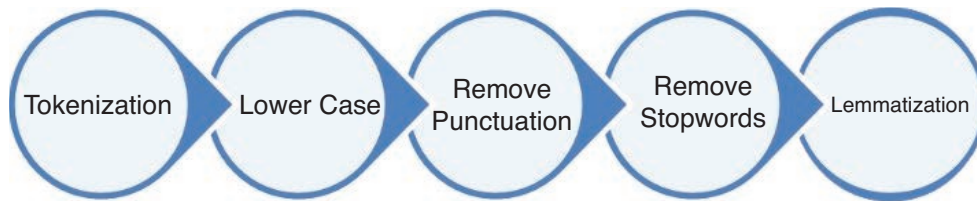


Figure 3. Data preprocessing step

3.2.1.1.1. Data Augmentation

In Figure 4, the data used in the study initially showed an imbalance between the categories, with '0' indicating people who were not depressed and '1' representing those who were. It was difficult to train and generalize the model effectively because of this unequal distribution.

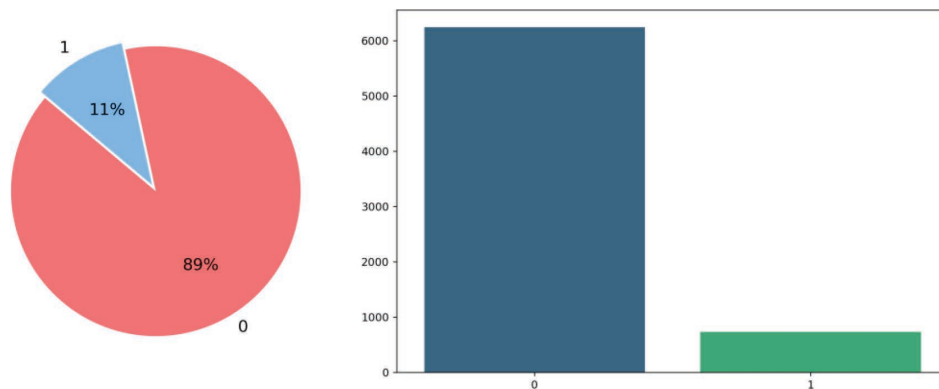


Figure 4. Graphical representation of imbalanced dataset

To solve the initial imbalance in our dataset and produce a balanced dataset, as shown in Figure 5, we used data augmentation techniques including synonym replacement and CharSwapAugmenter. By increasing data diversity and reducing the effects of imbalance, these techniques assisted in producing more variations of the text data, hence model robustness and performance improved.

3.2.1.2. Exploratory Data Analysis (EDA)

While performing the EDA, a special emphasis was put on the textual data and word clouds were created for the depressed and non-depressed classes separately (Figures 6 and 7). It must be mentioned

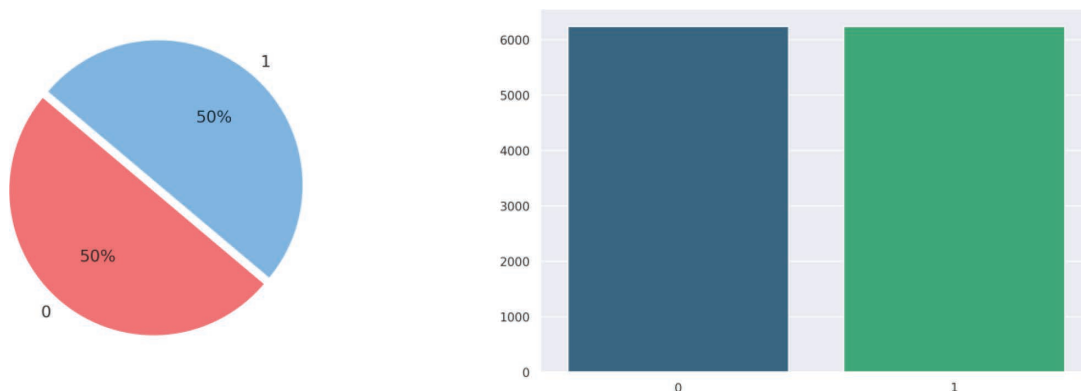


Figure 5. Graphical representation of balanced dataset



Figure 6. Word cloud for not depressed class



Figure 7. Word cloud for depressed class

that creating such word clouds allowed us to «capture» the most frequent and important words in each class. Thus, in a common word cloud for the depressed class, words associated with depressive feelings and thoughts were found, evidencing the common usage of such words among the members of this group. Similarly, the word cloud for the non-depressed class contained the phrases that were most commonly used by individuals who are not experiencing depression. Such visualizations made it possible to easily and rapidly discern the unique language features belonging to each class, enabling the identification of notable differences between the words used in the posts of depressed and non-depressed authors. This stage in the EDA was not only useful in gaining more insights from the data but also guided the later steps in terms of the analysis, including feature selection and model building.

Following the word cloud visualization, we further analyzed the frequency of words by plotting a bar chart. In the bar plots in Figures 8 and 9, the length of each bar corresponds to the frequency

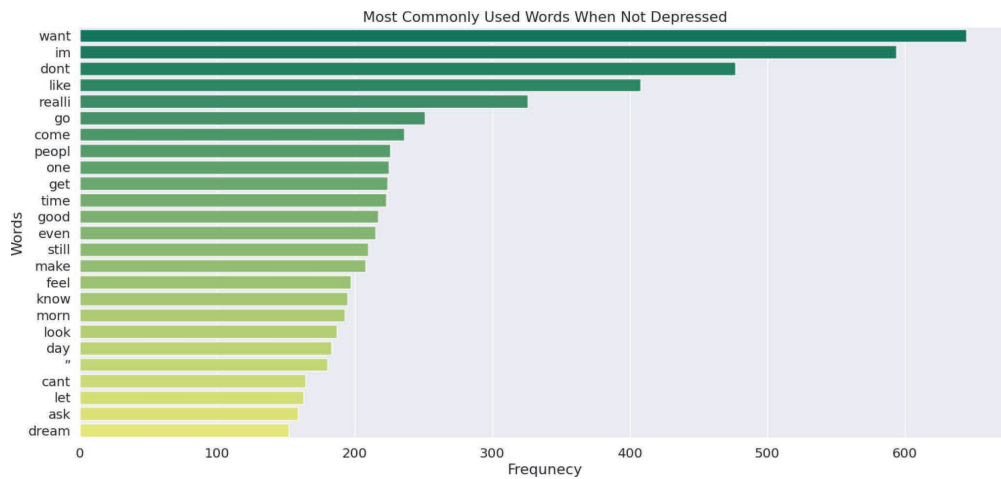


Figure 8. Most commonly used words in not depressed class

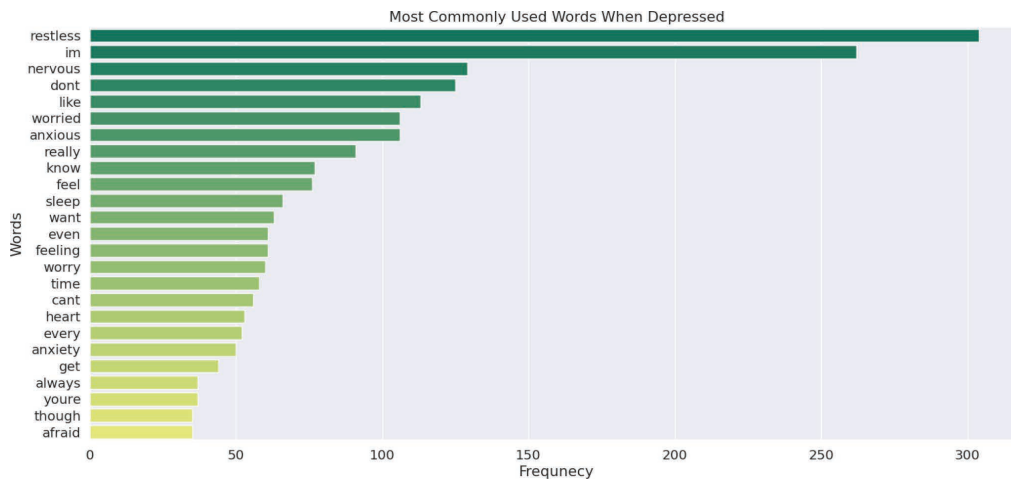


Figure 9. Most commonly used words in depressed class

of the respective word, providing a clear and quantifiable representation of word prominence in the dataset.

3.2.2. Machine Learning Techniques

Given the exponential growth of text data, manual classification becomes increasingly impractical. Therefore, it is essential to explore efficient methods for classifying large volumes of data quickly. Once classified, this data becomes valuable information, which can then be leveraged for strategic planning in business and industrial contexts. In this study, we implement two algorithms and conduct a comparative analysis to determine which one offers higher accuracy.

3.2.2.1. Multinomial Naïve Bayes

The Multinomial Naïve Bayes (MNB) classification algorithm is employed to classify discrete features, such as word frequencies in text classification (Lopez & Sumba, 2019). Although the multinomial distribution traditionally requires integer feature counts, fractional values such as those from TF-IDF can also be effective in practice (Usman et al., 2016). In the bag-of-words approach, each word is treated as a feature, with the order of words being irrelevant. Naïve Bayes relies on Bayes' rule of conditional probability (Bilal et al., 2016), and the MNB model is mathematically expressed by the equation:

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

Where A represents the hypothesis and B denotes the attribute in Figure 10.

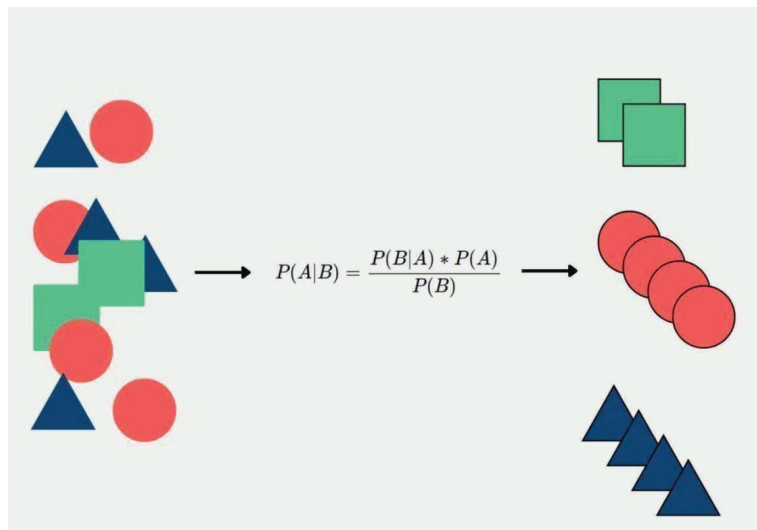


Figure 10. Multinomial Naive Bayes

3.2.2.2. Random Forest

There are numerous classification algorithms available, but Random Forest stands out as one of the top choices in machine learning (Figure 11). Although it can also be applied to regression tasks, its

primary use is for classification due to its versatility and straightforwardness. Random Forest enhances overall performance by combining multiple learning models (Kurnia et al., 2020; Zheng, 2019; Aggarwal et al., 2018). This technique involves aggregating many decision trees to form the forest, and achieving higher accuracy is possible when these trees are largely uncorrelated. Additionally, Random Forest can handle missing values effectively by imputing them (Matta & Saraf, 2020).

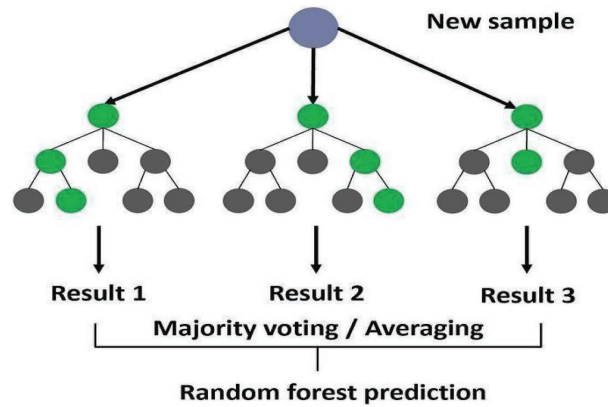


Figure 11. Random Forest

Decision tree classifiers are renowned for their exceptional performance, and since Random Forest is essentially an ensemble of decision trees, it inherits increased robustness and power. Even a single decision tree can deliver impressive results with high accuracy in classification problems (Charbuty & Abdulazeez, 2021). By leveraging multiple trees, Random Forest mitigates the limitations of individual decision trees, resulting in a more reliable and accurate classification model.

3.2.3. Explainable Machine Learning

While research into explainable machine learning (ML) is widely acknowledged as important, a unified understanding of explainability is still developing. From the field of explainable AI, we reference a definition from Montavon et al. (2018) «an explanation is the set of features within the interpretable domain that have contributed to a decision (e.g., classification or regression) for a given example».

3.2.3.1. LIME

Local Interpretable Model-agnostic Explanation (LIME) functions by utilizing a vector representation of documents. A widely used method for generating this representation is the Term Frequency-Inverse Document Frequency (TF-IDF) transformation (Luhn, 1957). The concept behind TF-IDF is relatively simple: it creates a vector of length D for each document. Given a dictionary ranging from w_1 to w_D , the j th component of this vector represents the significance of word w_j . This significance is calculated by multiplying two components: term frequency (TF), which measures how often the word appears in the document, and in-inverse document frequency (IDF), which indicates how rare the word is across the entire corpus. Essentially, a word's TF-IDF score will be higher if the word is common in the document but uncommon in the corpus as a whole. Consequently, frequently used words, such as «the» receive lower weights.

4. Results

To evaluate the effectiveness of our approach in classifying anxiety and depression, we implemented the Multinomial Naïve Bayes algorithm on the preprocessed data, achieving an accuracy of 92 %. The results include the confusion matrix and the Receiver Operating Characteristic (ROC) curve, which provide further insights into the classifier's performance by illustrating the true positive and false positive rates across different thresholds as shown in Figure 12.

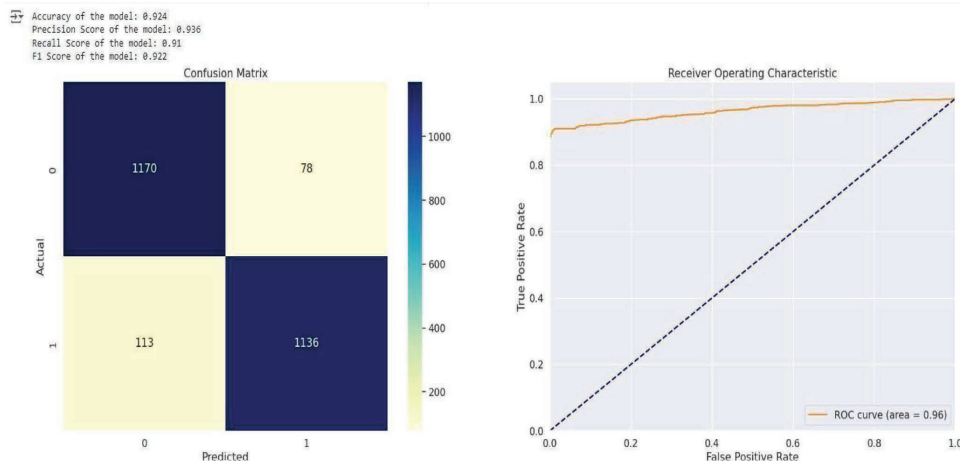


Figure 12. Naive Bayes insights

Following this, we implemented Random Forest which achieved an impressive accuracy of 99 %. To evaluate its performance, the confusion matrix and the Receiver Operating Characteristic (ROC) curve were examined, as seen in Figure 13. The confusion matrix gives us an extensive breakdown of the classifier's true positives, true negatives, false positives, and false negatives, allowing us to assess how well it performs across a range of categories. In contrast, the ROC curve illustrates the trade-offs between the true positive rate and false positive rate at various threshold values, which helps us understand how well the model can distinguish between classes. These criteria completely demonstrate the efficacy and robustness of the Random Forest model in relation to the previously evaluated model.

On the basis of the comparative analysis of both algorithms, as seen in Table 3, we concluded that the Random Forest outperformed the Multinomial Naïve Bayes in terms of accuracy and overall performance. The higher accuracy, along with the results from the confusion matrix and the ROC curve, indicate that the Random Forest model is more effective at correctly classifying the data.

To further understand the explainability of the Random Forest model, we applied an explainable machine learning approach using the LIME algorithm. LIME helps to interpret the model's predictions by identifying which features most significantly influenced the classifier's decisions, providing a more transparent view of how the Random Forest makes its predictions.

This representation pertains to the first instance, which is classified as «depressed». It highlights the 10 features that are most influential in the model's prediction for this instance. These features have been identified as significant contributors to the classifier's decision, providing insights into which aspects of the data are driving the model's determination of mental health in Figure 14.

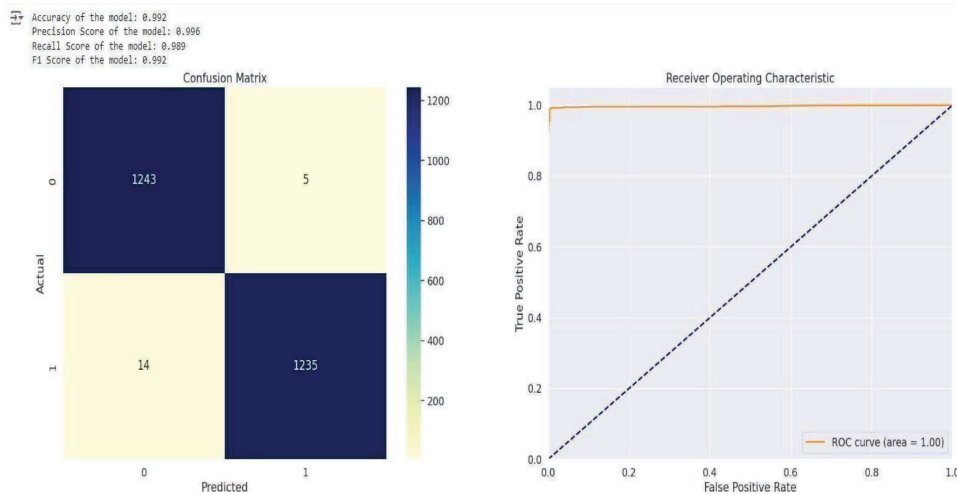


Figure 13. Random Forest insights

Table 3. Performance of experimental models

S.No	Embedding	Classifier	AUC (Area under curve)	Accuracy	Precision	F-1 Score
01	TF-IDF	Multinomial Naïve Bayes	0.96	92 %	0.93	0.92
02	TF-IDF	Random Forest	1	99 %	0.99	0.99

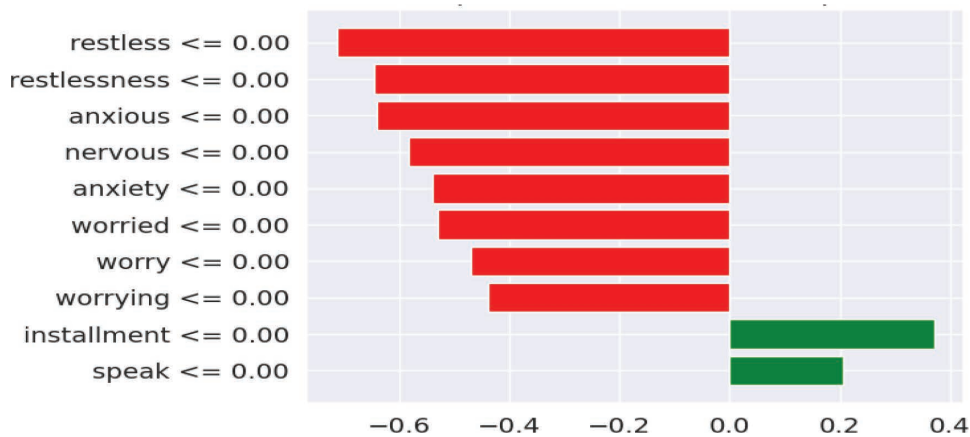


Figure 14. LIME representation

To further elaborate on the explainability of the Random Forest model, we demonstrated this using LIME by visualizing the explanations in a notebook graph, as seen in Figure 15. This graph illustrates the contributions of different features to the model's predictions, along with the probabilities assigned

to each class, indicating whether an individual is classified as depressed or not. This visualization provides a clear, interpretable insight into how the model makes its decisions for each instance.

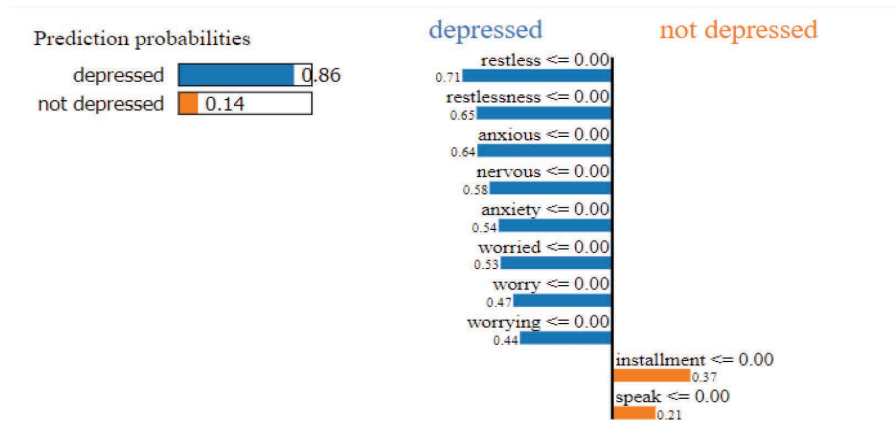


Figure 15. Probability of each class using LIME

5. Conclusion and Future Work

It is no wonder that mental health analysis and detection, especially regarding depression and its symptoms, has come to the forefront in natural language processing (NLP) research. In particular, this research explains how the depression signs may be recognized from different social media platforms based on machine learning techniques. The purpose of the study has been to identify the optimal machine learning approach to be used, through the application of different existing machine learning techniques and comparing their efficiency in this use case, namely, depression detection. The study was not limited to mere evaluation of performance, it also incorporated explainable machine learning using LIME, addressing machine learning's black-box nature and emphasizing the most important features.

It has been proven in the study that out of all the models used, the Random Forest model has the highest level of accuracy. It is further demonstrated that LIME is able to explain the relevance of the features used to make the prediction which makes it possible to ascertain and appreciate the Random Forest algorithm as one of the best in predicting mental health challenges such as depression.

As far as specific ideas are concerned, deep learning approaches could be integrated to enhance detection methods. In addition, extending the application to other explainable AI strategies in the field of NLP would probably improve both the accuracy and interpretability of models for the detection of mental health disorders. Such combination may offer fine tuning improvement, beyond prediction accuracy, also facilitating the mining for new information.

References

- Abdullah, D. M., & Abdulazeez, A. M. (2021). Machine learning applications based on SVM classification: A review. *Qubahan Academic Journal*, 1(2), 81–90.

- Aggarwal, A., Singh, J., & Gupta, D. K. (2018). A review of different text categorization techniques. *International Journal of Engineering and Technology (IJET)*, 7(3.8), 11–15.
- Al Hammadi, A. Y., Yeun, C. Y., Damiani, E., Yoo, P. D., Hu, J., Yeun, H. K., & Yim, M. S. (2021). Explainable artificial intelligence to evaluate industrial internal security using EEG signals in IoT framework. *Ad Hoc Networks*, 123, 102641. <https://doi.org/10.1016/j.adhoc.2021.102641>
- American Psychological Association. (2021, March 11). One year of unhealthy weight gains and increased drinking were reported by Americans coping with pandemic stress [Press release]. <http://www.apa.org/news/press/releases/2021/03/one-year-pandemic-stress>
- Becker, L. A., Penagos, H., Flores, F. J., Manoach, D. S., Wilson, M. A., & Varela, C. (2022). Eszopiclone and zolpidem produce opposite effects on hippocampal ripple density. *Frontiers in Pharmacology*, 12, 792148. <https://doi.org/10.3389/fphar.2021.792148>
- Bilal, M., Israr, H., Shahid, M., & Khan, A. (2016). Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree, and KNN classification techniques. *Journal of King Saud University - Computer and Information Sciences*, 28(3), 330–344. <https://doi.org/10.1016/j.jksuci.2015.10.003>
- Brailovskaia, J., Schillack, H., & Margraf, J. (2020). Tell me why are you using social media (SM)! Relationship between reasons for use of SM, SM flow, daily stress, depression, anxiety, and addictive SM use—An exploratory investigation of young adults in Germany. *Computers in Human Behavior*, 113, 106511. <https://doi.org/10.1016/j.chb.2020.106511>
- Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, 2(1), 20–28. <https://doi.org/10.38094/jastt20193>
- Elbadawi, M., Gaisford, S., & Basit, A. W. (2021). Advanced machine-learning techniques in drug discovery. *Drug Discovery Today*, 26(3), 769–777. <https://doi.org/10.1016/j.drudis.2020.12.005>
- Fabris, M. A., Marengo, D., Longobardi, C., & Settanni, M. (2020). Investigating the links between fear of missing out, social media addiction, and emotional symptoms in adolescence. *Addictive Behaviors*, 106, 106364. <https://doi.org/10.1016/j.addbeh.2020.106364>
- Gayathri, K., & Marimuthu, A. (2013, January). Text document pre-processing with the KNN for classification using the SVM. In *2013 7th International Conference on Intelligent Systems and Control (ISCO)* (pp. 453–457). IEEE. <https://doi.org/10.1109/ISCO.2013.6481162>
- Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: An integrative review. *Current Opinion in Behavioral Sciences*, 18, 43–49. <https://doi.org/10.1016/j.cobeha.2017.07.005>
- Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN model-based approach in classification. In *On The Move to Meaningful Internet Systems 2003* (pp. 986–996). Springer. https://doi.org/10.1007/978-3-540-39964-3_62
- Hu, C., Li, L., Li, Y., Wang, F., Hu, B., & Peng, Z. (2022). Explainable machine-learning model for prediction of in-hospital mortality in septic patients requiring intensive care unit readmission. *Infectious Diseases and Therapy*, 11(4), 1695–1713. <https://doi.org/10.1007/s40121-022-00508-9>
- Ibrahim, I., & Abdulazeez, A. (2021). The role of machine learning algorithms for diagnosing diseases. *Journal of Applied Science and Technology Trends*, 2(1), 10–19.
- Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. *WSEAS Transactions on Computers*, 4(8), 966–974.

- Joshi, G., Walambe, R., & Kotecha, K. (2021). A review on explainability in multimodal deep neural nets. *IEEE Access*, 9, 59800–59821. <https://doi.org/10.1109/ACCESS.2021.3070212>
- Kim, J., Lee, J., Park, E., & Han, J. (2020). A deep learning model for detecting mental illness from user content on social media. *Scientific Reports*, 10, 11846. <https://doi.org/10.1038/s41598-020-68886-0>
- Kurnia, R., Tangkuman, Y., & Girsang, A. (2020). Classification of user comment using word2vec and SVM classifier. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(1), 643–648.
- Lin, H., Jia, J., Guo, Q., Xue, Y., Li, Q., Huang, J., & Feng, L. (2014, November). User-level psychological stress detection from social media using deep neural network. In *Proceedings of the 22nd ACM International Conference on Multimedia* (pp. 507–516).
- Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., & Watkins, C. (2002). Text classification using string kernels. *Journal of Machine Learning Research*, 2(Feb), 419–444.
- Lopez, B., & Sumba, X. (2019). IMDb sentiment analysis.
- Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4), 309–317.
- Luo, X. (2021). Efficient English text classification using selected machine learning techniques. *Alexandria Engineering Journal*, 60(3), 3401–3409. <https://doi.org/10.1016/j.aej.2020.12.036>
- Matta, D. M., & Saraf, M. K. (2020). Prediction of COVID-19 using machine learning techniques.
- Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73, 1–15. <https://doi.org/10.1016/j.dsp.2017.10.011>
- Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. *Journal of Chemometrics*, 18(6), 275–285. <https://doi.org/10.1002/cem.889>
- Pirina, I., & Çöltekin, Ç. (2018, October). Identifying depression on Reddit: The effect of training data. In *Proceedings of the 2018 EMNLP Workshop SMM4H* (pp. 9–12).
- Praveen, S. V., Ittamalla, R., & Deepak, G. (2021). Analyzing Indian general public's perspective on anxiety, stress and trauma during Covid-19—a machine learning study of 840,000 tweets. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 15(3), 667–671. <https://doi.org/10.1016/j.dsx.2021.03.021>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). “Why should I trust you?” Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). <https://doi.org/10.1145/2939672.2939778>
- Rude, S., Gortner, E. M., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8), 1121–1133. <https://doi.org/10.1080/02699930441000030>
- Samek, W., & Müller, K. R. (2019). Towards explainable artificial intelligence. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K. R. Müller (Eds.), *Explainable AI: Interpreting, explaining and visualizing deep learning* (Lecture Notes in Computer Science, Vol. 11700, pp. 5–22). Springer. https://doi.org/10.1007/978-3-030-28954-6_1
- Sekulić, I., & Strube, M. (2020). Adapting deep learning methods for mental health prediction on social media. *arXiv preprint arXiv:2003.07634*. <https://doi.org/10.48550/arXiv.2003.07634>

- Sinha, P., & Sinha, P. (2015). Comparative study of chronic kidney disease prediction using KNN and SVM. *International Journal of Engineering Research*, 4(12), 608–612.
- Sljipecevic, D., Horst, F., Lapuschkin, S., Horsak, B., Raberger, A. M., Kranzl, A., & Zeppelzauer, M. (2021). Explaining machine learning models for clinical gait analysis. *ACM Transactions on Computing for Healthcare*, 3(2), 1–27. <https://doi.org/10.1145/3442387>
- Stirman, S. W., & Pennebaker, J. W. (2001). Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic Medicine*, 63(4), 517–522. <https://doi.org/10.1097/00006842-200107000-00001>
- Usman, M., Shafique, Z., Ayub, S., & Malik, K. (2016). Urdu text classification using majority voting. *International Journal of Advanced Computer Science and Applications*, 7(8). <https://doi.org/10.14569/IJACSA.2016.070823>
- Uysal, E., & Öztürk, A. (2018, May). Comparison of machine learning algorithms on different datasets. In *2018 26th Signal Processing and Communications Applications Conference (SIU)* (pp. 1–4). IEEE. <https://doi.org/10.1109/SIU.2018.8404331>
- Von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., & Schuecker, J. (2021). Informed machine learning—a taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Transactions on Knowledge and Data Engineering*, 35(1), 614–633. <https://doi.org/10.1109/TKDE.2021.3079836>
- Weld, D. S., & Bansal, G. (2019). The challenge of crafting intelligible intelligence. *Communications of the ACM*, 62(6), 70–79. <https://doi.org/10.1145/3282486>
- Yong, Z., Youwen, L., & Shixiong, X. (2009). An improved KNN text classification algorithm based on clustering. *Journal of Computers*, 4(3), 230–237. <https://doi.org/10.4304/jcp.4.3.230-237>
- Zheng, Y. (2019, November). An exploration on text classification with classical machine learning algorithm. In *2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)* (pp. 81–85). IEEE. <https://doi.org/10.1109/MLBDBI48998.2019.00021>
- Zinken, J., Zinken, K., Wilson, J. C., Butler, L., & Skinner, T. (2010). Analysis of syntax and word use to predict successful participation in guided self-help for anxiety and depression. *Psychiatry Research*, 179(2), 181–186. <https://doi.org/10.1016/j.psychres.2009.06.010>