Ajman University
College of Engineering and Information Technology
Department of Electrical and Computer Engineering

ELE466: Machine Learning

Project: Multi-Class Sentiment Classification of Mental Health Statements

Name: Bayan Alradi **ID: 202111426**

Course Learning Outcome	Mark Obtained	Maximum Mark
CLO		

Name: Aya Hmoud **ID: 202111446**

Course Learning Outcome	Mark Obtained	Maximum Mark
CLO		

Instructor: Dr. Mohamad Khairi Bin Ishak
Lab Instructor: Engr. Khaled Mahfouz

2nd May 2025

Abstract

This project explores applying Natural Language Processing (NLP) techniques paired with Machine Learning models to classify statuses of seven mental statuses based on text statements obtained from a variety of resources, such as posts on Twitter, Reddit, and other social media platforms. A preprocessing pipeline was developed, and several models were trained. Among them, the Random Forest classifier provided the highest testing accuracy, reaching 90%. This project emphasizes the huge role of machine learning in supporting mental analysis and prediction systems.

Objectives

- Build a classifier to detect the mental health status from a user statement.
- Compare performance across multiple machine learning models.
- Identify the best-performing approach and evaluate its effectiveness.
- Lay the groundwork for future integration into mental health chatbots and monitoring systems.

Introduction

Mental health is a vital pillar of the well-being and quality of life of a person. Depression, anxiety, and stress disorders are spreading due to the stressors in contemporary life, social isolation, and worldwide crises.About one in eight people in the world live with a mental disorder [1] . However, mental health disorder are often underdiagnosed and left untreated due to stigma and lack of access to care. The World Health Organization found that half the world's population lives in countries where there is just one psychiatrist to serve 200 000 or more people [1].

Social media has opened up new avenues for detecting the symptoms of mental health, especially on online forums and social media where individuals openly share their feeling and emotions through their posts. With Natural Language Processing , these text expressions are processed, and different machine learning models can be leveraged in developing intelligent systems that recognize various mental states to enable targeted support and early intervention. Such systems can contribute to mental health awareness, resource, allocation, and early prevention, especially in populations where access to care is limited.

Literature review

Over the last decade, work in mental health sentiment analysis has increased profoundly. Social media sites such as Reddit and Twitter have proven to be rich sources of data for analysing the psychological condition of individuals because user-generated content is rich and spontaneous. Text patterns have been found to be strongly correlated with emotional and mental health statuses, according to studies.

Early approaches primarily used basic models such as Naïve Bayes and Support Vector Machines with basic bag of words or TF-IDF features. In a study done in 2013 to predict depression from Twitter, an accuracy of 70% was achieved using SVC classifier [2]. These models were sufficient for early binary classification tasks such as depression or anxiety detection. More recent research has used deep learning models such as LSTMs, GRUs, and attention-based mechanisms to represent semantic nuances and content relationships in text.

More broadly, transformer models like BERT and RoBERTa have tuned out to excel particularly well on contextual meaning as well as affective subtlety detection. In a narrative review done by Zhang and others, they explored the power of LSTM and GRU networks which outperform traditional machine learning algorithms [3]. However, all of those are computationally heavy and require intensive annotated sets to fine-tune.

Our work builds on these bases with the application of classical and ensemble machine learning models atop a well-pre-processed TF-IDF feature space, as applied to a diverse and labelled mental health data set. It allows for a solid test of baseline methods before further development in future work using more sophisticated deep learning models.

Theoretical Background

In this section, the theoretical concepts behind Multinomial Naive Bayes (MNB) , Radial Basis Function Support Vector Machine, Logistic Regression, and Random Forest, the four machine learning models applied in this project will be discussed.

1) Multinomial Naive Bayes (MNB)

It is a probabilistic classification algorithm based on Bayes' Theorem, frequently used in tasks involving natural language processing such as spam detection, sentiment analysis and text classification [4]. The fundamental idea behind it is to assume the independency of all features from each other for each class label. In our case it means that the existence of a word in a statement does not affect the presence of another word. The formula for Bayes' theorem is given as:

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

Where:

- $P(A|B)$: This our goal which is the probability of event A happening given that event B has occurred.
- $P(B|A)$: The probability of event B happening when we know A has occurred.
- $P(A)$: The probability of event A happening without considering B. It's the "base rate" probability of A.
- $P(B)$: The probability of event B happening regardless of A.

In MNB, text frequency is used to estimate the probability of a sentence belonging to a specific class.We model the probability distribution of discrete characteristics using the Probability Mass Function (PMF). Given a class, PMF provides the probability of detecting a specific feature value. Let say X is a discrete random variable with range $R_X = \{x_1, x_2, x_3, \dots.\}$. The function $P_X(x_k) = P(X = x_k), for\ k = 1,2,3, \dots \dots.,$ is called the *probability mass function (PMF)* of X.

2) Radial Basis Function Support Vector Machine (RBF SVM)

RBF SVM is a supervised machine learning algorithm that can be applied for classification and regression problems, especially useful for non-linear and high-dimensional data [5]. In the early 1990s, Vladimir Naumovich Vapnik initially proposed the concept.

Given labelled data, the algorithm maps the data into a higher-dimensional feature space separating hyperplane which is a boundary that separates data into classes. It uses support vectors which are the most significant training or the nearest data points from each class and margins defined by these support vectors to maximize the distance between the hyperplane and the support vectors of each class.

It utilizes a kernel function called Radial Basis Function; to evaluate how similar the data points are in the space. It is defined as:

$$K(x, x') = e^{-gamma\|x - x'\|^2}$$

Where : x and x’ are input data points, and ||.|| is the Euclidean distance between them. Gamma is a hyperparameter that is responsible for the kernel’s width .

3) Logistic Regression

Logistic Regression is a supervised machine learning algorithm useful in classification tasks called predictive modelling where we are interested in the mathematical probability that a point belongs to a specific category or not. The main idea behind it is using a logistic (sigmoid) function. A sigmoid function is an S-shaped curve that turns any numerical value to a number in the range of 0 and 1. If the result of the function is greater than a defined threshold on the graph, the model predicts that the instance belongs to the class. If not, it predicts that the instance does not belong to that class.[6]

The role of Sigmoid Function in Logistic Regression is defined as:

$$P(y = 1) = \frac{1}{1 + e^{-\theta^T x}}$$

Where:

- $P(y=1)$ is the probability of the output being 1.
- θ represents the model's parameters (weights).
- x is the input feature.

This function transforms the linear combination $\theta^T x$ into a probability. If $\theta^T x$ is large and positive, the probability approaches 1, indicating a high confidence in class 1. Conversely, if $\theta^T x$ is large and negative, the probability approaches 0, indicating a high confidence in it.

4) Random Forest

Random Forest is a famous algorithm introduced by Leo Breiman and Adele Cutler, that produces a single output by combining the output of more than one decision tree as can be seen in figure 1. It is beneficial for non-linear classification problems.[7]

It basically searches for the best feature for splitting the node so that the resulting group is as distinguished as possible from the other groups and the members of each group are as similar as possible to each other.

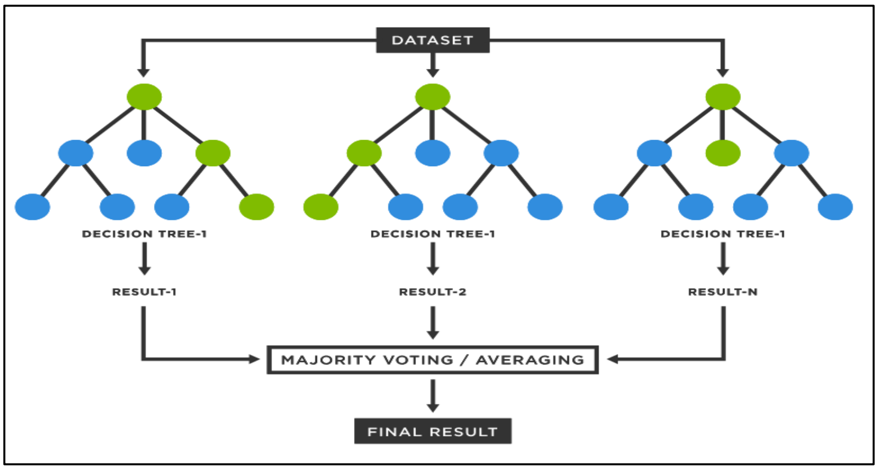


Figure 1 Architecture of Random Forest

Before training the model there are three main hyperparameters that must be set . These consist of node size, the number of trees, and the number of features sampled. Each tree in the collection of decision trees is composed of a bootstrap sample that is a data sample taken from the training set. A third of the bootstrap is set aside as test data and it is called out-of-bag sample. Feature bagging is used to introduce randomness , increasing the dataset variety and decreasing the corelation among all trees. For a classification problem the predicted class will be decided by a majority vote, which is the most common categorical variable. Lastly, the out-of-bag sample is used for cross-validation, finalizing the prediction.

Methodology

The methodology adopted in the project involves data preprocessing, feature extraction, data balancing, model training, evaluation, and prediction deployment for mental health text classification.

1) Dataset Overview

The dataset used in the project is called *Sentiment Analysis for Mental Health* available on Kaggle [7] . It is a compilation of 53,000+ cleaned and annotated statements tagged with one of seven labels: Normal, Depression, Suicidal, Anxiety, Stress, Bipolar, and Personality Disorder. The dataset was aggregated from Reddit, Twitter, and other platforms via multiple Kaggle mental health datasets. The csv. file includes the following 3 columns a *unique_id* column that identifies each entry, a *Statement* column which is the textual data, and lastly the *Mental Health Status* column compromised of the tagged mental health status each text.

2) Preprocessing

The preprocessing phase included several standard and customized raw text normalization steps. We defined a function called **preprocess_text** that involved expanding contractions using the contractions library to transform informal expressions (e.g., “don’t” to “do not”), which helps in capturing semantic meaning more effectively. Additionally, all words were lowercased to eliminate case sensitivity issues. Next, URLs and HTML tags were removed to eliminate irrelevant noise. After that, unnecessary characters are filtered away leaving simple punctuation and alphanumeric characters. Also, whitespace was normalized to preserve uniform spacing. Finally, Tokenization was performed on the clean text using **RegexpTokenizer**, which retained only meaningful words and punctuation.

To improve the quality of the data, we used a custom stopword list from **NLTK**, which preserved negations and first-person pronouns—crucial elements in psychological text analysis. Lemmatization was then applied using the **WordNetLemmatizer** to reduce words to their root forms.

3) Feature Extraction

TF-IDF vectorization was used to transform the cleaned text into numerical features. The TF-IDF (Term Frequency-Inverse Document Frequency) technique helps quantify the importance of a term in a document relative to the corpus. It is initialized with a maximum of 5000 features, which means that to represent the text data, 5000 of the most significant words are chosen based on their TF-IDF scores. The feature matrix **X_tfidf** is the input to the machine learning model. At the same time, the target labels (y), which stand for the relevant mental health categories for every text instance, are taken out of the status column.

Class imbalance was addressed using **RandomOverSampler** from the **imbalanced-learn library**, which duplicates examples in minority classes to balance the dataset. Then after balancing the dataset, the next step was to split the balanced dataset into training, validation and testing subsets. First, we split the dataset using a 60/40 split: 60% of the data was allocated to the training set, while the remaining 40% was held temporarily. Then, the temporary set was further split evenly into validation and test sets, resulting in 20% of the original data for validation and 20% for testing as can be seen in Table 1.

Train set size	68674
Validation set size	22891
Test set size	22892

Table 1 Data Partition Sizes After Balancing

4) Model Training

We evaluated four machine learning models all imported from the **scikit-learn library** based on their theoretical strengths in handling high-dimensional and sparse text data. Prior to predicting the labels of the validation set, each model was fitted to the training set. For comparison, the detailed classification reports and validation accuracy were calculated and saved.

To begin with, the Multinomial Naive Bayes (MNB) was chosen because it is well-suited for text-based classification tasks. However, its assumptions of the independency between features limits its overall performance in scenarios where complex relationships between words is important.

Then, Support Vector Machine (SVM) was chosen , specifically the linear variant due to its strength in handling high-dimensional data and its robustness to overfitting when regularization is applied. In text classification, where data is sparse and feature space is large, linear SVM is a commonly used baseline.

Logistic Regression's ability to model relationships between input features and output classes was a major reason for choosing it. It is useful when the classes are reasonably well-separated and there is a need for interpretability.

Lastly, Random Forest whose power lies in the capability of capturing complex patterns. Each model was trained using the TF-IDF features, followed by an 80/20 split into training and testing sets. A further validation set was used to fine-tune performance evaluation.

Results and Discussion

For the results obtained, we notice in Figure 2 that Random Forest outperformed scoring a 90% test accuracy due to its ability to capture non-linear patterns and handle high-dimensional TF-IDF data effectively. The ensemble approach improved generalization and robustness to noise, which is especially useful given the informal and varied nature of online mental health statements. While models like SVM and Logistic Regression performed well, but they were limited by their inability to capture deeper feature interactions.

Naive Bayes, although efficient, was less effective due to its strong independence assumptions, which are often violated in natural language. However, its speed and simplicity make it a good baseline for initial experimentation.

Further improvements can be made by introducing contextual embeddings such as BERT, which captures word meaning based on surrounding context rather than relying solely on frequency-based metrics.

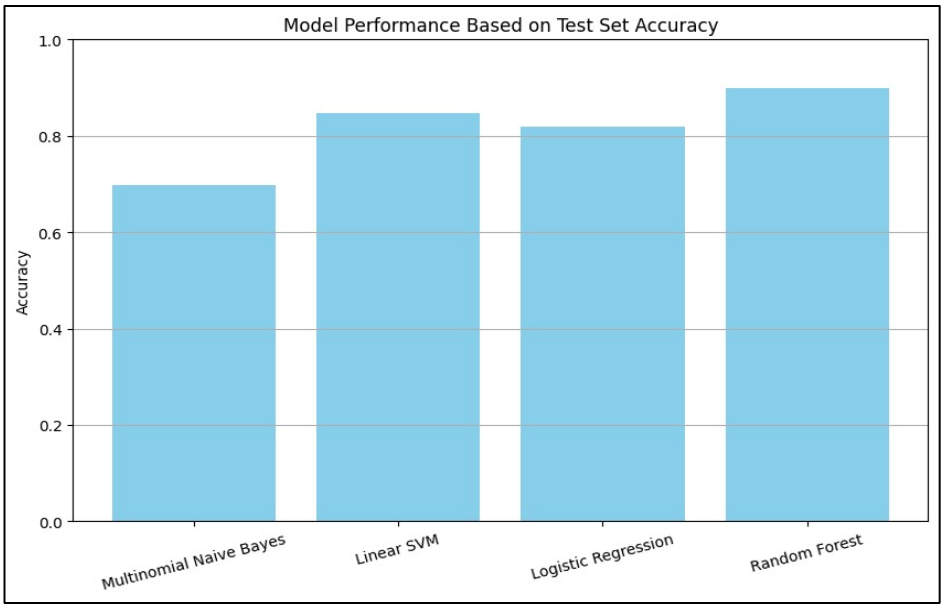
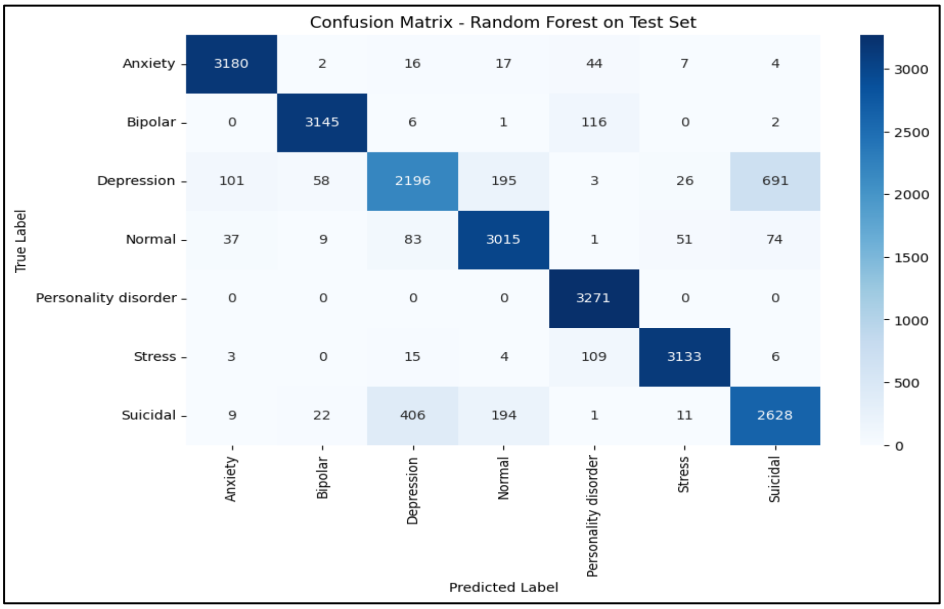


Figure 2 Test accuracies of each model

The confusion matrix in Figure 3 shows the performance of Random Forest on the test set. The model proves strong accuracy for most classes. However, there is a noticeable confusion between depression and suicidal and that is likely due to the semantic overlap between the two classes. There is a small part of stress which got misclassified as personality disorder which highlights feature similarity.



	Anxiety	Bipolar	Depression	Normal	Personality disorder	Stress	Suicidal
Anxiety	3180	2	16	17	44	7	4
Bipolar	0	3145	6	1	116	0	2
Depression	101	58	2196	195	3	26	691
Normal	37	9	83	3015	1	51	74
Personality disorder	0	0	0	0	3271	0	0
Stress	3	0	15	4	109	3133	6
Suicidal	9	22	406	194	1	11	2628

Figure 3 Random Forest confusion matrix

Conclusion

Through this project, we gained as a team valuable insight into the application of NLP with machine learning algorithms for mental health statuses classification. We learned how to clean and preprocess text data and apply TF-IDF to vectorize the data before feeding it into models like SVC with rbf kernel and Random Forest. After that we acquired new skills in understanding the results and analysing them to get digestible information that can be easily interpreted by people regardless their academic background which is truly beneficial in real-life settings. Overall, the project was not only an opportunity for us to apply the knowledge we gained throughout the semester but also gave us the chance to dig deep into the world of sentiment analysis and text classification.

Future directions for our project include but are not limited to:

- Fine-tuning transformer models (e.g., BERT, RoBERTa).
- Applying Named Entity Recognition (NER) for context enhancement.
- Building an interactive chatbot for real-time mental health support.
- Exploring emotion-aware attention networks for better contextual understanding.

References

[1] World Health Organization, World Mental Health Report: Transforming Mental Health for All, Geneva: World Health Organization, 2022.

[2] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting Depression via Social Media," in *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM),* Cambridge, MA, USA, Jul. 2013, pp. 128–137.

[3] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, "Natural language processing applied to mental illness detection: a narrative review," *npj Digital Medicine*, vol. 5, no. 46, Apr. 2022.

[4] H. A. H. Al-Shamri, "Bayesian Multinomial Naive Bayes Classifier to Text Classification," *International Journal of Computer Applications*, vol. 975, no. 8887, pp. 7–13, 2016.

[5] A. A. A. El-Sappagh, M. Elmogy, and S. R. M. El-Bakry, "A comprehensive biomedical text mining framework for extracting and analyzing unstructured biomedical data," *Computers in Biology and Medicine*, vol. 134, p. 104481, 2021.

[6] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd ed., draft chapter 5, Jan. 12, 2025.

[7] TIBCO Software Inc., "What is a Random Forest?," Spotfire Glossary, 2025. [Online]. Available: https://www.spotfire.com/glossary/what-is-a-random-forest

[8] S. Sarkar, "Sentiment Analysis for Mental Health," Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health