Detection of Mental Health Disorder using Text Corpora and Machine Learning Approach

Kulkarni Swapna
School of Technology, Sub Campus
S.R.T.M.U.N, Peth, Latur

Agnihotri P.P.
School of Technology, Sub Campus,
S.R.T.M.U.N, Peth, Latur

ABSTRACT
Psychological disorder is a caused due to continuous feeling of sadness and not a single hope from outside world. This feeling negatively effects on mental and emotional wellbeing of patients. Psychological disorders are a serious and growing global health concern [1]. Psychological disorders patients have become a leading key player to the global health crisis**.** Traditional diagnostic approaches are often time-consuming and can be affected by factors such as patient unwilling to self-report symptoms owing to social shame. In today's many people undergone psychological disorder so it is essential to detect the psychological disorder.

In this paper The Machine learning (ML) algorithm was used to get a potential solution by using digital data for early detection and interventions of psychological disorders. This paper also focuses development of Naïve Bayes model for classification mental health disorders. In this model a text-based data corpora have been used which was collected from the Kaggle and a Naïve Bayes (NB) algorithm is also used for the classification of mental health disorders. This model gives 88% accuracy to detect the mental health disorders like Anxiety, and Depression.

Keywords
Machine Learning, Health Disorders, Mental Illness, Text based corpora

1. INTRODUCTION
Depressions is one of the mental illnesses [3] that interfere with a person's everyday emotions, thoughts and behaviour as well as their overall health.[1]. The prevalence of psychological disorders necessitates the development of improved diagnostic tools [6]. NLP techniques applied to text-based data sources, such as clinical notes, and interviews collected from Kaggle database has high probability of detection of mental illness [8]. However, text analysis alone might miss subtle behavioural, so vocal patterns which are indicative of mental health issues as well as multifaceted approaches that integrate different data types offer a more comprehensive understanding of an individual's mental state.

This paper investigates the application of a machine learning Naïve Bayes algorithm which uses textual features and potentially other complementary data derived from text (e.g., temporal features and sentiment scores) to detect psychological disorders. The theoretical principles of Naïve Bayes classifiers handle text data for improving diagnostic accuracy of detection of psychological disorder.

2. LITERATURE REVIEW
i. Psychological Disorders and Text-Based Detection

Psychological disorders include many conditions that impact a person's thoughts, emotions and behaviours [3]. Anxiety, depression, bipolar disorder, and schizophrenia are common psychiatric disorders. Natural language processing (NLP) techniques, such as sentiment analysis, emotion detection, and topic modelling, are increasingly used to gain insights from text data for mental health applications [5]. Research indicates an upward trend in NLP-driven mental illness detection.

ii. Naive Bayes Algorithm

Naïve Bayes is a supervised machine learning algorithm that is often used for text classification tasks [10]. It is based on Bayes' theorem and assumes that features are independent, which simplifies the classification problems.

Multinomial Naive Bayes is suitable for text classification, where the features are word frequencies or counts.

iii. Multichannel Data Fusion

Multichannel data fusion combines information from different data streams to achieve a more comprehensive understanding or improve system performance. In psychological disorder detection, this can combine text features, behavioural cues extracted from text (e.g., frequency of posting, time of day), or sentiment analysis scores [2].

Different fusion approaches exist, including early, late, and hybrid fusions. Early fusion combines raw data prior to processing. Late fusion process modalities are used independently before merging the outputs for decision-making.

3. METHODOLOGY

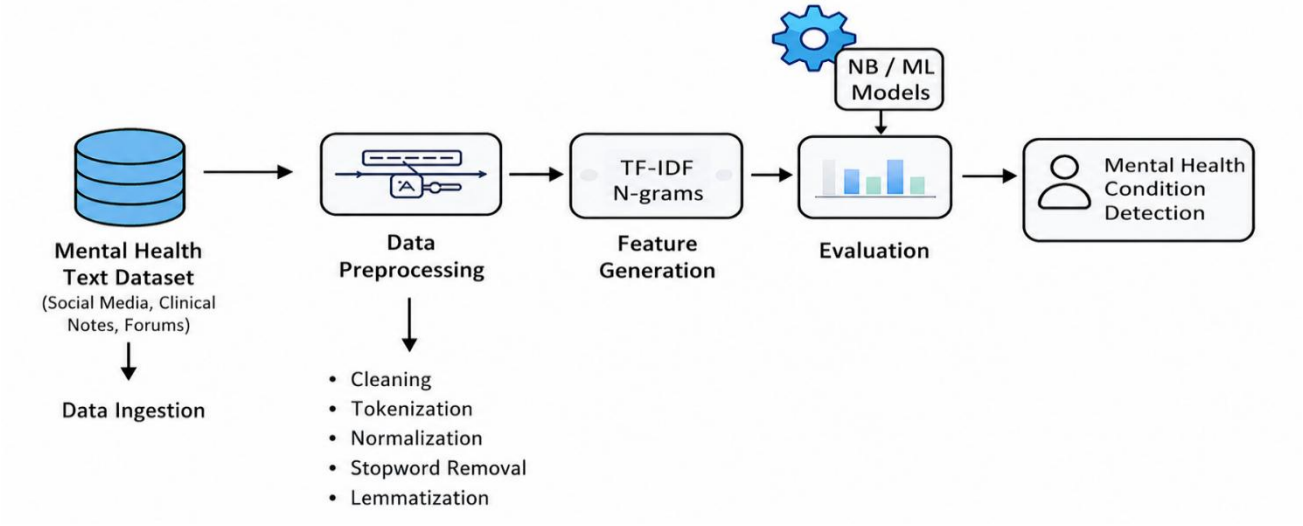

Fig 1: Navie Byes Model

The proposed model for detecting mental health disorders from text corpora is designed as a modular and scalable architecture that processes raw textual input and classifies it into predefined categories such as Healthy, Anxiety, and Depression. The system integrates natural language processing (NLP), feature generation, and a Naïve Bayes classifier to achieve accurate and efficient disorder prediction. The complete workflow consists of five major components: data acquisition, data pre-processing, feature extraction, Feature generation, and classification with evaluation.

3.1 Data Acquisition Layer

The system begins with the collection of a labelled text dataset containing samples annotated with corresponding psychological disorder categories. In this study Kaggle dataset has been used for experiment. Each document is paired with a class label—typically Healthy (0), Anxiety (1), or Depression (2)—which is essential for supervised learning. Ethical considerations regarding privacy, confidentiality, and data security are strictly observed during this stage.

3.2 Text Pre-processing Layer

Since raw text contains noise and variability, the system applies several pre-processing steps to standardize input data before feature extraction. These steps follow established NLP methodologies [9].

Pre-processing operations include:

- Lowercasing: All characters are converted to lowercase to maintain uniformity.
- Noise Removal: Numbers, special characters, URLs, and irrelevant symbols are removed.
- Tokenization: Text is segmented into individual words (tokens).
- Stopword Removal: Common non-informative words such as “the,” “is,” and “and” are filtered out.
- Lemmatization/Stemming: Words are reduced to their base or root form to minimize vocabulary size and enhance generalization.

The output of this module is a cleaned and standardized textual dataset suitable for feature extraction.

3.3 Feature Extraction

This layer transforms the pre-processed text into numerical representations that can be processed by the machine-learning model. Multiple feature types are extracted to capture linguistic, semantic, and behavioural patterns relevant to mental health detection.

TF-IDF Features:

Term Frequency–Inverse Document Frequency (TF-IDF) is used to quantify the importance of words within a document relative to the entire corpus. This representation is particularly effective for identifying key psychological cues and high-impact terms.

Linguistic and Stylistic Features:

These features include part-of-speech tags, usage patterns of personal pronouns, sentence structure cues, and stylistic markers that may indicate emotional or cognitive states.

Sentiment and Emotion Features:

Sentiment analysis tools assign polarity scores (positive, negative, neutral), while emotion lexicons detect psychological signals such as sadness, anger, fear, or hopelessness—common indicators of anxiety or depression.

Temporal and Behavioural Features:

If timestamp data is available, temporal patterns such as posting frequency, timing, and irregular activity patterns are extracted, providing additional behavioural context to improve prediction accuracy.

3.3.1 Feature generation

To fully utilize the diverse feature types, the system supports a multifaceted feature fusion strategy.

- Early Fusion

All extracted features—TF-IDF vectors, linguistic features, sentiment scores, and temporal attributes—are concatenated into a single composite feature vector, which is then used to train a unified Naïve Bayes classifier.

- Late Fusion

Separate Naïve Bayes classifiers are trained independently on different feature groups. Their outputs are later combined using

majority voting or weighted averaging to generate the final prediction.

- Hybrid Fusion

This strategy merges the strengths of both early and late fusion by performing early fusion on tightly related features and employing late fusion for decision-level integration across independent modalities.

The flexibility of these fusion approaches allows the model to adapt to varying dataset characteristics and improves diagnostic robustness.

3.4 Classification Layer: Multinomial Naïve Bayes Model

The core of the system is a Multinomial Naïve Bayes classifier, well suited for text-based classification tasks. The model uses the following principles:

- Prior Probability: Computed based on the distribution of each class in the training dataset.
- Likelihood Estimation: Calculated for each word given a class label using frequency information.
- Laplace Smoothing: Applied to avoid zero-probability issues for unseen words.
- Posterior Probability: Derived using Bayes' theorem to assign the most likely class to each input document.

This classifier is computationally efficient, interpretable, and effective for high-dimensional text data.

3.5 Evaluation and Output Layer

After training, the model is evaluated on the test set using multiple performance metrics, including accuracy, precision, recall, F1-score, and AUC-ROC.

The system achieved an overall accuracy of 88%, with strong performance on *Healthy* and *Anxiety* classes. However, performance on the *Depression* class was limited due to insufficient sample size, highlighting the importance of balanced datasets in psychological disorder detection.

3.5.1 Experiments and Evaluation

The effectiveness of the proposed multifaceted Naïve Bayes framework was examined through a series of experiments designed to assess its capability in identifying mental health disorders using text data. The evaluation followed recognized practices used in prior mental-health classification research [7], employing standard performance metrics such as accuracy, precision, recall, F1-score, and AUC-ROC.

3.5.2 Experimental Setup

The cleaned and feature-engineered dataset—consisting of TF-IDF vectors, linguistic features, and sentiment attributes—was divided into training and testing subsets using a 70:30 split. The Multinomial Naïve Bayes classifier was trained on the fused feature set under the multifaceted framework. Evaluation was conducted on the reserved test data to ensure unbiased performance measurement.

4. ALGORITHM FOR PREDICTION OF MENTAL HEALTH USING TEXT CLASSIFICATION

This algorithm deals with the prediction of status of mental health of the patients with the help of text corpora and Naive Bayes method. This algorithm incorporates seven steps which is shown as follows

Input: A dataset of text documents D, where each document $di$is labelled with a classCk

(e.g.,$C1$='Disorder Present', C_2='DisorderAbsent').
Output: A trained Naïve Bayes classification model capable of predicting the class of new, text data.

Phase 1: Data Preparation and Feature Engineering

1. Initialize Data and Labels:

- Acquire the labelled dataset $D$of text documents.

2. Preprocessing Function Preprocess (Text Data):

For every data element in dataset (D) do for loop until

D != EOF //For each document diin$D$:

Li ← lower(di) //Change all text to lowercase.

Li ←sub (Li) //Clear with numbers, and special characters.

Li ←Split(Li) //Tokenize the text into individual words (tokens).

Li ←remove_stpowords(Li) //Remove common stop words (e.g."a","is") using a standard lexicon.

Li ←lemmatize_or_stem(Li) //Apply lemmatization or Temmingto reduce words to their base form.

End For

Return D_cleaned //**Return** the cleaned and tokenized dataset $Dcleaned$.

3. Feature Extraction (Vectorization):

Create a vocabulary V from all unique tokens in D_cleaned //Use the cleaned dataset $Dcleaned$to create a vocabulary of all unique words.

Apply a **Bag-of-Words (BoW)** or **TF-IDF** vectorizer to convert each text document into a numerical feature vector.

Result:

X→ document -term matrix

y→ corresponding class labels

// A numerical data matrix $X$where rows represent documents and columns represent word features (counts or TF-IDF scores), and a corresponding label vector y.

4. Data Splitting Function Split Data (X, y):

training set: ($Xtrain$, $Ytrain$)

testing set:($Xtest$, $Ytest$).

// (Common split: 70% train, 30% test).

Return the four data subsets.

Phase 2: Model Training (Multinomial Naïve Bayes)

5. Model Training Function Train Model (X_train, y_train):

- Initialize a Multinomial Naïve Bayes classifier.
- Calculate the **Prior Probabilities**(Ck)for each
- $P(C_k) = \frac{\text{Number of documents in class} C_k}{\text{Total number of training documents}}$
- Calculate the **Likelihoods**$P(wj|Ck)$ for each word feature $w_j$given a class C_k
- Use **Laplace Smoothing** to prevent zero probabilities for words that do
- Not appear in a specific class during training.
- $P(w_j|C_k) = \frac{\text{Count of} w_j \text{in class} C_k + 1}{\text{Total words in} C_k + \text{Total unique words in vocabulary}}$
- The model stores these learned prior probabilities and likelihoods.
- **Return** the trained NB model.

Phase 3: Prediction and Evaluation

6. Prediction Function Predict (Trained Model, X_test):

For each test document d_{test}inX_{test}:

$$P(C_k|d_{test}) \propto P(C_k) \times \prod_{j=1}^{m} P(w_j|C_k)$$

The algorithm assumes feature independence (the "naïve" assumption).

- Assign the class with the highest posterior probability as the final prediction for the document.
- The algorithm assumes feature independence (the name assumption)

7. Evaluation Function Evaluate (y_test, predictions):

Compare the predicted labels ($y_{predictions}$) with the actual true l[po abels (y_{test}).

Calculate performance metrics:

Accuracy

Precision

Recall

F1-Score

5. RESULTS

The performance of prediction of mental health using text is measured by implementing the Naive Bayes classification algorithm. The table 1 shows the result obtained for the subjects those who have mental health disorder.

Table 1: performance evaluation of Naive Bayes algorithm

Class Label	Precision	Recall	F1-Score	Support
Healthy (0)	0.94	0.82	0.88	975
Anxiety (1)	0.84	0.94	0.89	956
Depression (2)	0.00	0.00	0.00	2
Accuracy			**0.88**	1933
Macro Avg	0.59	0.59	0.59	1933
Weighted Avg	0.89	0.88	0.88	1933

The model gives an overall accuracy of **88%**, and **94 %** for the Healthy and Anxiety categories. However, the Depression class showed insufficient predictive performance due to dataset has the subjects those who are facing mental health issues they are at very beginning stage.

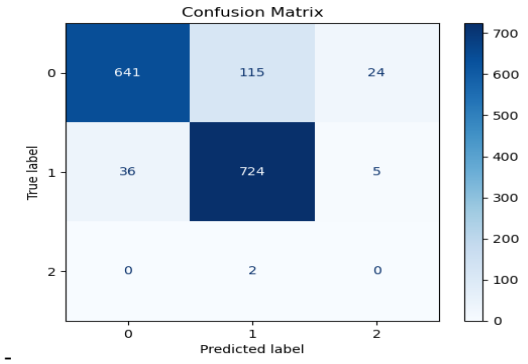

Fig 2: Confusion Matrix

In Fig 2 the confusion matrix indicates that the model achieves strong performance on the predominant classes. It correctly classifies **641 instances of class 0** and **724** instances **of class 1**, demonstrating effective learning and clear discrimination between these categories.

Comparative Evaluation

To verify the value of the multifaceted feature fusion approach, performance was compared against baseline systems such as:

- **Unimodal Naïve Bayes** trained solely on TF-IDF text features
- **Traditional machine-learning models** including Support Vector Machines (SVM) and Random Forests
- Table 2: Comparison of performance of various algorithms

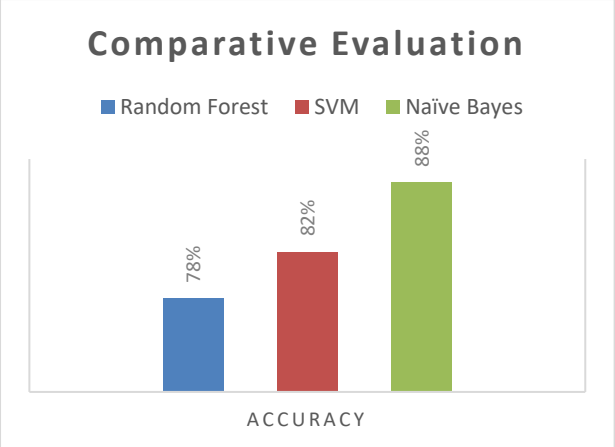

Fig 3: Comparative Evolution

The Fig 3 shows that the SVM and Random Forest algorithms were also implemented using same dataset for prediction of mental health disorder. The results obtained from SVM and Random Forest algorithms are 82% and 78% respectively.

The results indicate that the multifaceted Naïve Bayes classifier provides improved overall accuracy **88%** for mental health detection as compared to SVM and Random Forest algorithm.

Ethical considerations, including patient privacy, data security, algorithmic bias, and the necessity for clinical validation, are paramount. The model's output should be considered advisory and requires validation by mental health professionals.

6. CONCLUSION

This study aims to demonstrate the effectiveness of a multifaceted Naïve Bayes algorithm for detecting psychological disorders from text-based data. By integrating diverse information derived from textual inputs and leveraging the strengths of Naïve Bayes classifiers, the goal is to develop a model that contributes to improved accuracy, early detection, and more proactive mental health interventions.

Although the naive bayes algorithm gives high probability to predict the mental state of the subjects using text corpora, but it may possible text corpora does not guarantees us for correct answers given by the subjects, so in future, this study includes the audio as well as EEG signals of the subject for detection or prediction of mental state of each subjects so that the guaranteed or concrete results will help medical professional to identify the level of mental disorder.

7. REFERENCES

[1] Merino, M et al. Body perceptions and psychological well - being: A review of the impact of social media and physical measurements on self- esteem and mental health with a focus on body image satisfaction and its relationship with cultural and gender factors. Healthcare 12 (14),1396(2024)

[2] Atrey, P. K., Hossain, M. A., El Saddik, A., & Kankan Halli, M. S. (2010). Multimodal fusion for multimedia analysis: A survey. Multimedia Systems, 16(6), 345–379.

[3] American Psychiatric Association. (2022). Diagnostic and Statistical Manual of Mental Disorders (5th ed., text revision). APA Publishing.

[4] Bruckner, T., Li, Y., & Kumar, S. (2022). Ethical considerations in AI-based mental health systems. Journal of Digital Ethics, 4(1), 11–21.

[5] Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review. IEEE Transactions on Affective Computing, 1(1), 18–37.

[6] Calvo, R. A., & Milne, D. (2020). Natural language processing in mental health applications. Annual Review of Clinical Psychology, 16, 79–110.

[7] Garg, S., Verma, P., & Singh, R. (2021). Machine learning techniques for depression detection using text. International Journal of Computer Applications, 174(32), 1–6.

[8] Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: A review. Current Opinion in Behavioural Sciences, 18, 43–49.

[9] Jurafsky, D., & Martin, J. H. (2021). Speech and Language Processing (3rd ed.).

[10] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.

[11] Merino, M., et al. (2024). Body perceptions and psychological well-being… Healthcare, 12(14), 1396.

[12] World Health Organization. (2022). World Mental Health Report: Transforming Mental Health for All.