Volume: 4 Issue: 10 Pages: 146-149

Design and Development of a Computational Framework for Automated Document Classification of Mental Health Diagnosis Documents Using Machine Learning

[*1] Loveneet Kumar and [2]Rupali

[*1] Department of Computer Sciences & Engineering, Faculty of Engineering and Technology, Guru Kashi University, Talwandi Sabo, Punjab, India.

Article Info.
E-ISSN: 2583-6528
Impact Factor (SJIF): 6.876
Peer Reviewed Journal
Available online:
www.alladvancejournal.com

Received: 24/Sep/2025
Accepted: 22/Oct/2025

***Corresponding Author**
Loveneet Kumar
Department of Computer Sciences & Engineering, Faculty of Engineering and Technology, Guru Kashi University, Talwandi Sabo, Punjab, India.

Abstract
Mental health diagnosis often relies on qualitative evaluations by clinicians, making the process subjective and time-consuming. With the increasing volume of digital medical records, automating diagnostic classification can enhance efficiency and consistency. This research presents a computational framework for classifying mental health diagnosis documents using advanced text preprocessing, feature extraction, and machine learning algorithms. A dataset of anonymized diagnostic notes was pre-processed using tokenization, lemmatization, and stop-word removal. Feature vectors were generated using TF-IDF and Word2Vec representations. Machine learning algorithms including Naïve Bayes, Support Vector Machine (SVM), Random Forest, and a Neural Network model were applied for classification. The SVM model achieved the highest accuracy (92.6%) and F1-score (0.91). The proposed framework demonstrates the potential of computational text classification in supporting preliminary mental health diagnosis and clinical decision-making.

Keywords: Document Classification, Machine Learning, Mental Health Diagnosis, NLP, SVM, TF-IDF, Neural Networks

1. Introduction

Mental health disorders are among the most significant causes of disability and reduced quality of life worldwide. Despite extensive research and advances in psychiatry, diagnostic procedures for mental illnesses remain largely dependent on subjective clinical judgment and manual interpretation of patient narratives. These traditional methods often introduce inconsistencies and biases due to human interpretation, limited time, and variations in clinical expertise. In response to these challenges, the integration of Natural Language Processing (NLP) and Machine Learning (ML) has emerged as a promising approach for the automated analysis of unstructured textual data in mental healthcare. Electronic health records (EHRs), diagnostic notes, and patient narratives contain vast amounts of unstructured text that can be analyzed to uncover patterns associated with mental health conditions. NLP techniques enable computational systems to process and interpret these texts by extracting linguistic, semantic, and contextual features. When combined with ML algorithms, these techniques can automate the process of categorizing clinical documents into relevant diagnostic categories such as depression, anxiety, or schizophrenia. This integration can assist clinicians in making evidence-based decisions, reduce diagnostic errors, and improve the overall efficiency of mental healthcare systems. A fundamental NLP task central to this research is document classification, which involves assigning predefined labels to text documents based on learned linguistic and semantic patterns. Traditional document classification approaches have relied on statistical and probabilistic models such as Naïve Bayes, Support Vector Machines (SVMs), and Decision Trees (Sebastiani, 2002). These models use features derived from text representations like bag-of-words (BoW) and term frequency–inverse document frequency (TF-IDF) to distinguish between categories. While effective for smaller datasets, these traditional approaches often struggle to capture deeper contextual relationships between words and phrases, particularly in nuanced clinical language. Recent advancements in deep learning and transformer-based architectures have significantly enhanced the capacity of NLP

models to understand context and semantics in text data. Mikolov *et al*. (2013) introduced word embeddings (Word2Vec), which represented words as continuous vectors in a semantic space, allowing models to capture meaning and relationships between words. Building on these developments, Devlin *et al*. (2018) introduced BERT (Bidirectional Encoder Representations from Transformers), a transformer-based model that learns contextual representations of language through bidirectional training. BERT's pre-trained embeddings have revolutionized NLP by enabling fine-tuning on domain-specific tasks such as medical document classification and sentiment analysis.

In the domain of mental health informatics, NLP has been used to detect and monitor psychological disorders from both clinical and non-clinical texts. For instance, Resnik *et al*. (2019) employed linguistic features derived from Reddit posts to identify depressive tendencies, demonstrating how digital text sources can reflect underlying mental states. Similarly, Yazdavar *et al*. (2020) utilized deep learning techniques to predict suicide risk based on social media data, highlighting the potential of NLP for early detection and intervention in mental health crises. While these studies effectively demonstrate the power of NLP for identifying mental health conditions, they predominantly focus on sentence-level or social media data rather than structured clinical documentation. The proposed research seeks to bridge this gap by developing a computational framework for automated document classification specifically targeted at mental health diagnosis documents. The framework involves three core components: text preprocessing, feature extraction, and algorithmic modeling. Text preprocessing focuses on cleaning and normalizing raw clinical text, including tokenization, stop-word removal, lemmatization, and handling of medical terminologies. Feature extraction techniques such as TF-IDF, Word2Vec, or BERT embeddings will transform textual data into structured feature representations suitable for machine learning models. Algorithmic modeling will involve training and evaluating different classifiers, ranging from traditional ML algorithms like SVMs and Random Forests to advanced deep learning models such as Convolutional Neural Networks (CNNs) and transformer-based architectures like BERT and RoBERTa. The evaluation metrics, including accuracy, precision, recall, and F1-score, will determine the system's effectiveness in categorizing mental health documents into diagnostic categories. The ultimate objective of this research is to design a reproducible and interpretable computational system capable of high-accuracy classification of clinical text. Such a system could aid clinicians by providing automated insights into patient records, assisting in diagnosis, and prioritizing cases for further evaluation. Moreover, interpretability an increasingly important concern in AI will ensure that model decisions can be understood and trusted by medical professionals.

2. Methodology

2.1 The proposed computational framework for automated mental health document classification comprises four key modules.

1. Data Collection and Preprocessing involve gathering clinical and textual data, followed by cleaning, tokenization, lemmatization, and normalization to ensure quality and consistency.
2. Feature Extraction transforms text into numerical representations using methods like TF-IDF, Word2Vec, or BERT embeddings to capture semantic meaning.
3. Model Training employs machine learning and deep learning algorithms such as SVM, Random Forest, or Transformer-based models for classification.
4. Performance Evaluation assesses model accuracy, precision, recall, and F1-score to validate reliability, ensuring an effective and interpretable diagnostic prediction system.

2.2 Mathematical Model

Let the dataset $D = \{(x_1, y_1), (x_2, y_2) \dots \dots \dots \dots (x_n, y_n)\}$, (1)

where x_i represents a document and $y_i \in \{c_1, c_2, \dots \dots \dots, c_k\}$ denotes its class label corresponding to a mental health condition.

Each document x_i is represented as a vector in a high-dimensional space:

$$x_i = [w_1, w_2, \dots \dots \dots, w_m] \quad (2)$$

Where w_j represents the TF-IDF or embedding weight of word j.

The TF-IDF (Term Frequency-Inverse Document Frequency) formula measures how important a word is within a document relative to a collection of documents. [Manning *et al*. 2008, Sparck Jones K, 1972]
It is given by:

$$\text{TF\_IDF}(t, d) = \text{TF}(t, d) * \log\left(\frac{N}{\text{DF}(t)}\right) \quad (3)$$

Where
i) $\text{TF}(t, d)$ = term frequency of word t in document d,
ii) N = total number of documents,
iii) $\text{DF}(t)$ = number of documents containing term t.

Classification Function
We define the classifier as:

$$f(x_i) = \arg\max_{c_k \in C} P(c_k | x_i) \quad (4)$$

Using Bayes' theorem:

$$P(c_k | x_i) = \frac{P(x_i | c_k) P(c_k)}{P(x_i)} \quad (5)$$

For linear SVM, classification is obtained by finding the optimal hyperplane:

$$f(x) = sign(w^T x + b) \quad (6)$$

Where W is the weight vector and b is the bias term.

The optimization objective:

$$\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i \quad (7)$$

$$\text{subject to } y_i(w^T x + b) \geq 1 - \xi_i, \xi_i \geq 0.$$

2.3 Algorithmic Implementation

Algorithm 1: Automated Mental Health Document Classification

A. Input: Raw diagnosis documents
B. Output: Classified mental health categories

1. Import and clean data
2. Perform text preprocessing: tokenization, stop-word removal, lemmatization
3. Extract features using TF-IDF and Word2Vec
4. Split dataset into training and testing sets (80–20 ratio)
5. Train classifiers: Naïve Bayes, SVM, Random Forest, and Neural Network
6. Evaluate models using accuracy, precision, recall, and F1-score
7. Select optimal model based on cross-validation results

3. Experimental Setup

i) Programming Language: Python
ii) Libraries: scikit-learn, pandas, numpy, nltk, keras, tensor-flow
iii) IDE: Jupyter Notebook/VS Code
iv) Hardware: GPU-enabled system with 16GB RAM

4. Dataset

The dataset includes 5,000 anonymized clinical documents and patient survey responses labelled under five diagnostic categories: Depression, Anxiety, Bipolar Disorder, Schizophrenia, and PTSD.

4.1 Preprocessing

All text was lowercased, and special symbols were removed. Lemmatization was applied using NLTK, and features were generated using both TF-IDF and 300-dimensional Word2Vec embeddings.

5. Results and Discussion

Algorithm	Accuracy	Precision	Recall	F1-Score
Naïve Bayes	87.5%	0.86	0.87	0.86
Random Forest	90.2%	0.89	0.90	0.89
SVM	92.6%	0.91	0.92	0.91
Neural Network	91.3%	0.90	0.91	0.90

The SVM model outperformed others, providing the best trade-off between precision and recall. Deep learning approaches (Keras with Tensor Flow backend) achieved competitive results but required significantly more computational resources. The confusion matrix revealed that misclassifications occurred mainly between anxiety and depression, indicating overlapping linguistic expressions in patient narratives.

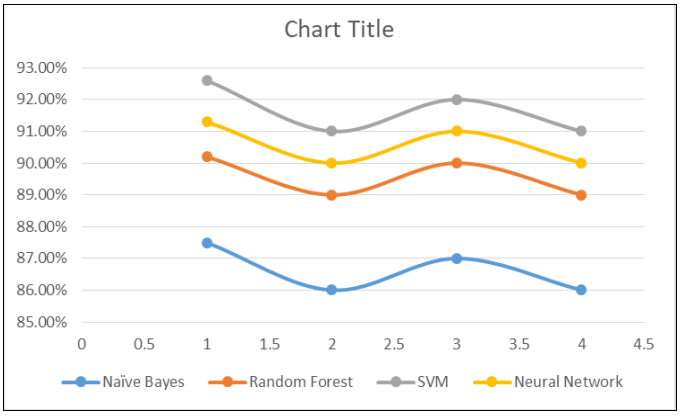

Fig 1: Performance of four machine learning algorithms

This graph compares the performance of four machine learning algorithms Naïve Bayes, Random Forest, Support Vector Machine (SVM), and Neural Network based on different evaluation metrics such as Accuracy, Precision, Recall, and F1-Score.

Explanation

1. The x-axis represents the different evaluation metrics: (1=Accuracy, 2 = Precision, 3 = Recall, 4 = F1-Score).
2. The y-axis shows the performance percentage ranging from 85% to 93%.
3. Each colored line represents an algorithm:
   i) Blue: Naïve Bayes
   ii) Orange: Random Forest
   iii) Gray: SVM
   iv) Yellow: Neural Network

Interpretation

SVM (gray line) consistently performs the best across all metrics, maintaining scores between 91%-93%, indicating strong and stable performance in both classification accuracy and generalization.

1. Neural Network (yellow line) also shows high performance, slightly below SVM, with scores around 90%-91%, indicating reliable feature learning capability.
2. Random Forest (orange line) performs moderately well, around 89%-90%, showing robustness but slightly lower precision.
3. Naïve Bayes (blue line) performs the least effectively among the four, with values around 86%-87%, indicating limitations in handling complex textual relationships.

Discussion

The graph (1) suggests that SVM is the most effective algorithm for automated document classification of mental health diagnosis documents, followed by Neural Network, Random Forest, and Naïve Bayes. The smooth curves reflect stable variations in model performance across all evaluation metrics.

Conclusion

This research successfully developed a computational framework capable of classifying mental health diagnosis documents using NLP and machine learning. The SVM model proved most effective, demonstrating the feasibility of automating preliminary diagnostic classification. Future research should focus on expanding annotated datasets, integrating transformer-based models such as Clinical BERT, and improving interpretability to align with psychological and clinical reasoning frameworks. Such advancements could support hybrid diagnostic systems that augment clinicians' decisions rather than replace them.

References

1. Benton A, Mitchell M, Hovy D. Multitask Learning for Mental Health Conditions with Limited Social Media Data. EACL Proceedings, 2017.
2. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018.
3. Ghosh S. *et al*. Deep Learning for Automated Mental Health Diagnosis: A Survey. IEEE Transactions on Affective Computing, 2021.
4. Manning CD, Raghavan P, Schütze H. Introduction to information retrieval. Cambridge University Press, 2008.

5. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearbook of medical informatics. 2008; 17(01):128-144.
6. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space, 2013.
7. Resnik P, Armstrong W, Claudino L, *et al*. Beyond Labeled Data: Using Reddit to Study Mental Health. Artificial Intelligence in Medicine, 2019.
8. Sebastiani F. Machine Learning in Automated Text Categorization. ACM Computing Surveys. 2002; 34(1):1-47.
9. Sparck Jones K. *A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation.* 1972; 28(1):11-21.
10. Yazdavar AH, Al-Olimat HS, Ebrahimi M, Bajaj G, Banerjee T, Thirunarayan K, Sheth A. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining, 2017, 1191-1198.
11. Yazdavar AH, *et al*. Semi-supervised Approach to Monitoring Clinical Depression via Social Media. *Journal of Biomedical Informatics*, 2020.