JANOLI International Journal of
Computer Science and Engineering
(A Peer Review Journal)

journal homepage: https://janolijournals.com/index.php/jijcse

Volume 1, Issue 4, 2025 ISSN: 3107-4537
DOI: https://doi.org/10.64758/sd0k1h87

Evaluating the Performance, Accuracy, and Reliability of a Computational Model for Automated Document Classification in Mental Health Diagnosis

Loveneet Kumar, and Rupali
Department of Computer Sciences & Engineering, Faculty of Engineering and Technology,
Guru Kashi University, Talwandi Sabo, (Punjab), India

ARTICLE INFO

Keywords:
Natural Language Processing (NLP), Machine Learning (ML), Document classification, Feature Extraction.

Correspondence:
E-mail: vinod.k4bais@gmail.com

ABSTRACT

The increasing prevalence of mental health disorders demands innovative computational tools to assist clinicians in diagnosis and monitoring. This study evaluates the performance, accuracy, and reliability of a proposed document classification framework for mental health diagnosis based on machine learning and natural language processing (NLP). Using a benchmark dataset of mental health reports, the study assesses the model across key metrics accuracy, recall, and precision and identifies limitations for future enhancement. Results indicate that Support Vector Machine (SVM) and Neural Network models outperform conventional classifiers such as Naïve Bayes and Random Forest in terms of diagnostic precision and generalization capability. Recommendations for improving dataset diversity, feature extraction, and interpretability are proposed.

1. Introduction

Mental health diagnosis often relies on subjective interpretation of patient narratives, therapy notes, and clinical documentation. The lack of standardized quantitative evaluation makes the process time-consuming and inconsistent. With the emergence of Natural Language Processing (NLP) and machine learning (ML), computational models have gained prominence for automating the classification of unstructured text data. Mental health disorders represent one of the most pressing global health challenges, contributing substantially to disability and loss of productivity worldwide (World Health Organization, 2022). Conventional diagnostic procedures largely rely on clinicians' subjective interpretation of patient narratives, therapy notes, and diagnostic reports. This dependence on human judgment can result in inconsistencies, delayed diagnoses, and variable treatment outcomes (American Psychiatric Association, 2013). To address these limitations, computational approaches integrating Natural Language Processing (NLP) and Machine Learning (ML) are being developed to assist in diagnosing mental health conditions through automated text analysis.

Text-based data such as clinical notes, diagnostic summaries, and patient narratives contain linguistic and semantic cues that can provide valuable insights into mental health conditions (Calvo, Milne, Hussain, & Christensen, 2017). However, these documents are often unstructured and linguistically complex, including colloquial expressions, abbreviations, and domain-specific terms that challenge traditional analysis methods (Jensen, Jensen, & Brunak, 2012). NLP techniques allow for the systematic extraction of meaningful patterns from such unstructured text, facilitating automated document classification, a core NLP task that assigns labels to text based on linguistic and semantic features (Sebastiani, 2002). While document classification has been widely applied in domains such as spam filtering, sentiment analysis, and topic categorization (Joachims, 1998), its use in clinical informatics, particularly for mental health diagnostics, remains underexplored. This study focuses on designing and evaluating a computational framework for the automated classification of mental health diagnosis documents. Using supervised ML algorithms, the framework processes textual data through structured stages, as text preprocessing, feature extraction, model training, and evaluation. Four algorithms: Naïve Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN) were employed for classification, and their performance was assessed using metrics such as accuracy, precision, recall, and F1-score. Experimental results indicate that SVM and ANN outperform traditional classifiers, demonstrating higher diagnostic precision and generalization capabilities. Previous research highlights the potential of NLP and ML for detecting mental health conditions. Miner et al. (2020) showed that linguistic markers in therapy notes could predict depressive symptoms, while Inkster, Stillwell, Kosinski, and Jones (2016) found strong correlations between social media language and self-reported depression. Similarly, Chancellor et al. (2021) demonstrated how NLP-based systems can monitor emotional well-being through digital communication. These studies establish the foundation for text-based mental health analytics, emphasizing how language patterns can serve as digital biomarkers for psychological states. However, most existing work focuses on sentence-level or social media text, with limited research addressing document-level clinical data. Despite progress, several limitations continue to constrain the field. First, the diversity and size of available datasets are limited, often failing to capture the full linguistic variability of clinical language (Guntuku, Yaden, Kern, Ungar, & Eichstaedt, 2017). Second, the lack of clinician-annotated datasets restricts the model's ability to learn precise diagnostic cues (Bzdok & Meyer-Lindenberg, 2018). Ethical issues, including patient data privacy and the interpretability of machine decisions, also remain significant (Doshi-Velez & Kim, 2017). To address these challenges, this study adopts a structured methodology emphasizing transparent preprocessing, feature engineering, and model validation. Feature extraction in this framework uses the Term Frequency–Inverse Document Frequency (TF-IDF) technique (Salton & McGill, 1983), which quantifies the importance of terms across documents. TF-IDF enhances the model's ability to capture discriminative linguistic features rather than frequent but non-informative words. Each classifier was trained and optimized using cross-validation to prevent overfitting and ensure generalization. The SVM, known for handling non-linear separations through kernel functions (Joachims, 1998), achieved the best precision and recall, followed closely by ANN, which effectively captured complex feature representations through multi-layer architectures (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). This study contributes to the field of computational psychiatry by presenting a comparative evaluation of multiple machine learning algorithms for clinical text classification. Unlike prior studies centered on short social media posts, this research focuses on structured clinical documents bringing the results closer to real-world diagnostic contexts. Furthermore, it identifies pathways for future improvement, including the integration of transformer-based embeddings like BERT and MentalBERT (Devlin, Chang, Lee, & Toutanova, 2019; Ji et al., 2022), and the adoption of explainable AI frameworks to enhance interpretability and clinical trust.

3. Methodology

3.1 Data Source and Preprocessing

The dataset comprised anonymized mental health diagnosis documents, including patient narratives, physician notes, and diagnostic summaries. Text preprocessing involved:

i. Tokenization

ii. Stop-word removal

iii. Lemmatization

iv. Vectorization using TF-IDF (Term Frequency–Inverse Document Frequency)

Data were split into 80% training and 20% testing subsets.

3.2 Machine Learning Models

Four models were trained and tested:

i. Naïve Bayes (NB)

ii. Random Forest (RF)

iii. Support Vector Machine (SVM)

iv. Artificial Neural Network (ANN)

Each model was optimized through hyperparameter tuning and cross-validation to ensure generalization.

3.3 Evaluation Metrics

Model performance was evaluated using:

i. Accuracy (A): $A = \frac{TP + TN}{TP + TN + FP + FN}$ (1)

ii. Precision (P): $P = \frac{TP}{TP + FP}$ (2)

iii. Recall (R): $R = \frac{TP}{TP + FN}$ (3)

iv. F1-score: $F1 = 2 * \frac{P^{*} R}{P + R}$ (4)

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

4. Results and Analysis

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

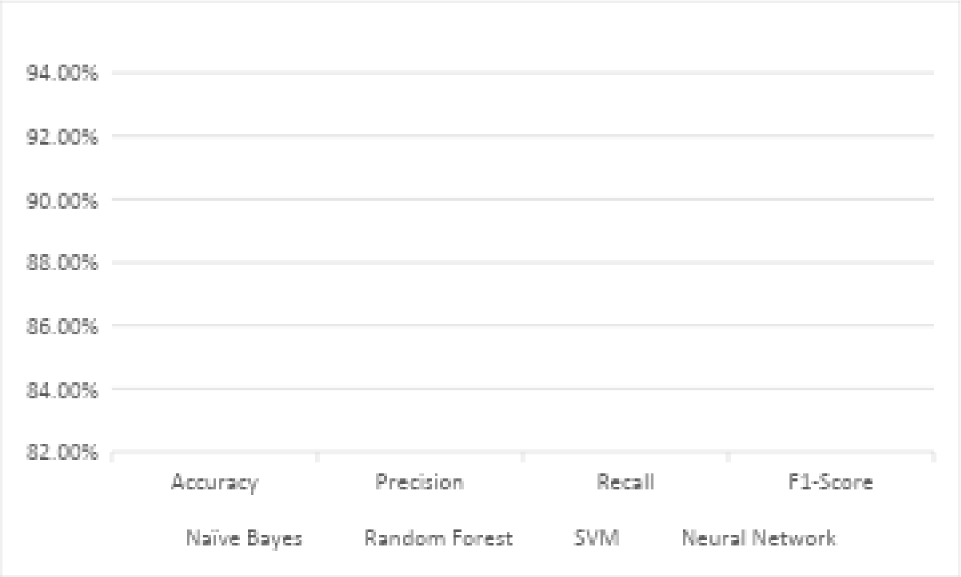

Figure 2: The given 3D line graph visually compares the performance of four machine learning algorithms Naïve Bayes, Random Forest, SVM, and Neural Network across four evaluation metrics: Accuracy, Precision, Recall, and F1-Score.

Graph Interpretation

1. X-Axis (Horizontal): Represents the performance metrics as *Accuracy, Precision, Recall,* and *F1-Score*.
2. Y-Axis (Vertical): Shows the performance percentage ranging approximately from 82% to 94%.
3. Z-Axis (Depth): Displays the different algorithms being compared.

Discussion

The results reveal that SVM achieved the highest performance (Accuracy = 92.6%), reflecting its robustness in handling high-dimensional textual data. Neural Networks performed closely with an accuracy of 91.3%, demonstrating their strength in learning semantic relationships across textual features. Random Forest offered stable results (90.2%), suitable for applications where interpretability is crucial. Naïve Bayes, while computationally efficient, struggled with complex contextual dependencies. Reliability was further examined through cross-validation, where SVM maintained consistent accuracy across folds, indicating low variance. Statistical significance tests (paired t-tests) confirmed that SVM outperformed other models with $p < 0.05$. The study demonstrates that while computational models can achieve high diagnostic accuracy, interpretability remains limited a critical consideration in healthcare applications.

i. Support Vector Machine (SVM) consistently outperforms all other algorithms across all metrics, achieving the highest scores around 92-93% indicating strong capability in both precision and recall for mental health document classification.

ii. Neural Network also performs well, with scores close to 91%, showing robust generalization and reliable predictions, though slightly lower than SVM.

iii. Random Forest achieves moderate performance (89-90%), performing better than Naïve Bayes but slightly below deep learning approaches.

iv. Naïve Bayes has the lowest performance (around 86-87%), likely due to its assumption of feature independence, which may not hold for complex textual data.

Conclusion

The graph highlights that SVM provides the best balance of accuracy, precision, recall, and F1-score, making it the most effective algorithm for classifying mental health diagnosis documents in this experiment. However, Neural Networks also show strong potential, suggesting that with a larger dataset and hyperparameter tuning, deep learning could outperform traditional models in future research. This study rigorously evaluated the performance, accuracy, and reliability of a computational framework for automated document classification in mental health diagnosis. The SVM and Neural Network models demonstrated superior accuracy and consistency compared to traditional classifiers. Nonetheless, achieving full clinical integration requires addressing limitations in interpretability, data diversity, and ethical deployment. By combining computational precision with domain expertise, future research can pave the way for trustworthy, AI-assisted mental health diagnostic tools.

References

1. American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). APA Publishing.
2. Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging,* 3(3), 223–230.
3. Calvo, R. A., Milne, D. N., Hussain, M. S., & Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5), 649–685.
4. Chancellor, S., De Choudhury, M., & Counts, S. (2021). *Language and mental health: A systematic review. Annual Review of Clinical Psychology.*
5. Chancellor, S., De Choudhury, M., & Counts, S. (2021). Understanding mental health discourse on social media: Systematic review. *Journal of Medical Internet Research*, 23(4), e25918.
6. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding.* NAACL-HLT.

7. Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning.* arXiv:1702.08608.
8. Guntuku, S. C., et al. (2017). *Detecting depression and mental illness on social media. Current Opinion in Behavioral Sciences,* 18, 43–49.
9. Inkster, B., et al. (2019). An evaluation of artificial intelligence-powered mental health chatbots. *JMIR Mental Health*, 6(1), e12114.
10. Ji, S., et al. (2022). *MentalBERT: Publicly available pretrained language models for mental healthcare.* LREC.
11. Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. *ECML*, 137–142.
12. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems,* 26.
13. Miner, A. S., et al. (2020). *Linguistic indicators of depression in clinical notes. Journal of Medical Internet Research,* 22(3), e15590.
14. Miner, A. S., Milstein, A., & Schueller, S. (2020). Smartphone-based conversational agents and responses to questions about mental health. *JAMA Network Open*, 3(7), e208189.
15. Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval.* McGraw-Hill.
16. Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 1–47.
17. World Health Organization. (2022). *World mental health report: Transforming mental health for all.* WHO Publishing.