IBIMA Publishing
Communications of International Proceedings
https://ibimapublishing.com/p-articles/45AI/2025/4541325/
Vol. 2025 (20), Article ID 4541325

Natural Language Processing in Healthcare: A Case Study on Depression Detection*

Mudasir Ahmad WANI[1], Kashish Ara SHAKIL[2], Talal Abdulmohsen S ALDHABAAN[3], Muhammad Asim[1], Ogerta ELEZAJ4 and Mohammed ELAFFENDI[1]

[1]EIAS Data Science Lab, College of Computer and Information Sciences, Prince Sultan University, Riyadh, Saudi Arabia
[2]Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia
[3] College of Computer and Information Sciences, Prince Sultan University, Riyadh, Saudi Arabia
[4]University of Birmingham, Birmingham, United Kingdom
School of Computing and Digital Technology

Correspondence should be addressed to: Mudasir Ahmad WANI, mwani@psu.edu.sa

* Presented at the 45th IBIMA International Conference, 25-26 June 2025, Cordoba, Spain

Abstract

With the rapid growth of electronic health records (EHRs), clinical notes, and physician summaries, healthcare systems are generating vast amounts of unstructured textual data. Unlocking meaningful insights from this information especially to support early detection of mental health conditions like depression remains a significant challenge. While Natural Language Processing (NLP) offers powerful tools to address this, there's still a need to explore its effectiveness in real-world clinical contexts.

In this study, we apply a range of NLP techniques to clinical text to detect early signs of depression. Our pipeline includes domain-specific preprocessing steps like tokenization, lemmatization, and lexical normalization. We use TF-IDF and contextual embeddings for feature extraction, followed by classification using traditional models (Logistic Regression, Random Forest) and deep learning approaches (LSTM, ClinicalBERT).

We obtained promising results, Logistic Regression and LSTM models achieved perfect ROC-AUC scores of 1.000, with F1-scores of 0.800, reflecting strong balance between precision and recall. ClinicalBERT achieved high precision (1.000) but struggled with recall (0.400), resulting in a lower F1-score of 0.571. Random Forest, by contrast, performed poorly across most metrics. These findings show the potential of combining classic and modern NLP methods for early depression detection and suggest that even simpler models can deliver strong results with well-engineered features. We hope this work supports further efforts in building intelligent, interpretable clinical decision-support tools in mental health care.

Keywords: Natural Language Processing (NLP), Medical Text Analysis, Deep Learning, Artificial Intelligence, Clinical BERT.

Cite this Article as: Mudasir Ahmad WANI, Kashish Ara SHAKIL, Talal Abdulmohsen S ALDHABAAN, Muhammad Asim, Ogerta ELEZAJ and Mohammed ELAFFENDI, Vol. 2025 (20) "Natural Language Processing in Healthcare: A Case Study on Depression Detection " Communications of International Proceedings, Vol. 2025 (20), Article ID 4541325, https://doi.org/10.5171/2025.4541325

Introduction

Electronic Medical Records (EMRs) and Electronic Health Records (EHRs) have become fundamental components of modern healthcare infrastructure, offering structured and semi-structured digital formats to record patient data. These systems improve population health by understanding the level of diseases and wellbeing of individuals. These systems enhance healthcare delivery by ensuring timely access to medical information, reducing administrative costs, and supporting secure and confidential communication of patient data [1] . However, a substantial portion of clinical data in EMRs/EHRs remains unstructured, including physician notes, discharge summaries, pathology reports, and radiology interpretations.

This unstructured textual data poses significant challenges for information retrieval and decision-making due to its complexity, variability, and lack of standardization. If managed properly they have potential to improve quality of patients health [2]. Natural Language Processing (NLP), a branch of Artificial Intelligence (AI), has emerged as a vital tool to extract actionable insights from clinical narratives. By leveraging linguistic, statistical, and deep learning approaches, NLP enables automated processing of clinical text for tasks such as named entity recognition, concept normalization, relation extraction, and text classification. A similar approach had been adopted by authors in [3] where an open-source NLP platform has been developed which extracts information from clinical text. When combined with machine learning techniques, especially deep neural networks, and transformers such as Deep Bidirectional Transformers (BERT), NLP can facilitate scalable solutions for biomedical information extraction, predictive analytics, and clinical decision support systems (CDSS) [4]. Furthermore, NLP techniques if applied on clinical notes can help in identification of life-threatening consequences such as critical limb ischemia [2].

Biomedical NLP is inherently interdisciplinary, integrating methodologies from computational linguistics, machine learning, and bioinformatics. It focuses on extracting information from clinical texts, with one of its core objectives being transformation of free-text medical data into structured representations to enhance EHR usability and inform CDSS [5]. Despite its promise, the domain faces several issues, particularly making secondary use of data for information extraction to automatically encode clinical information from raw text samples. They lead to challenges which include the prevalence of domain-specific abbreviations, syntactic and semantic ambiguities, de-identification requirements, and the use of non-standard clinical terminologies across different healthcare providers [6]. Historically, clinical text analysis and information extraction from clinical data was performed manually, which limited scalability and incurred high labor and temporal costs. The automation of this process through NLP using information extraction tasks such as text classification, relationship extraction and named entity recognition not only improves efficiency but also supports early detection of diseases and proactive patient care [7], [8]. For instance, NLP-driven systems can aid in identifying early symptoms of chronic conditions, adverse drug events, and patient risk stratification.

This paper aims to explore the evolving role of NLP in the healthcare domain by critically reviewing key methodologies and highlighting emerging applications. In doing so, it underscores the immense potential of NLP to convert unstructured clinical narratives into structured, actionable knowledge that can enhance diagnostic accuracy and treatment efficacy, assist clinicians in making timely and informed decisions, and enable data-driven healthcare policy formulation and resource allocation.

Furthermore, this study presents a detailed case study demonstrating how NLP can be practically applied to detect early signs of depression from clinical notes.

Main Contributions

The main contributions of the paper are as under

In-depth Analysis of the Role of NLP in the Healthcare Domain:

We provide a comprehensive and technically grounded exploration of how NLP is being increasingly integrated into healthcare workflows to process and interpret unstructured clinical text. This contribution outlines the foundational and advanced methodologies that allow NLP systems to generate structured, semantically rich representations of clinical narratives.

Case Study: Depression Detection from Clinical Notes

We present a realistic and domain-specific case study focusing on early detection of depressive symptoms through clinical narratives. This case study simulates the NLP pipeline on a corpus of clinical notes and provides empirical insights into how NLP tools can be tailored to extract mental health indicators such as suicidal ideation and fatigue from unstructured tex.

Implementation of Machine Learning and Transformer-Based Models

We implement benchmark NLP models ranging from traditional algorithms to modern deep learning approaches, with a focus on clinical text classification. including Logistic Regression, Random Forests using hand-crafted features such as TF-IDF and transformer architectures such as ClinicalBERT.

Performance Evaluation of the models

We evaluated model performance using standard metrics such a accuracy, precision, recall, F1-score, and AUC-ROC to determine their efficacy in classifying clinical narratives as depressive or non-depressive. This allows for a robust, quantitative assessment of NLP's role in clinical text mining and predictive modeling.

Overview of NLP

A language is a mechanism for communication comprising of written or verbal means, which can be words or a set of symbols. A language which is developed, spoken or written by humans for communication is called as a natural language for example English, Arabic, Hindi, Japanese and French. NLP is a branch of computer science which deals with processing of this natural language. It describes techniques such as syntax parsing and Chomsky transformation methods [9].

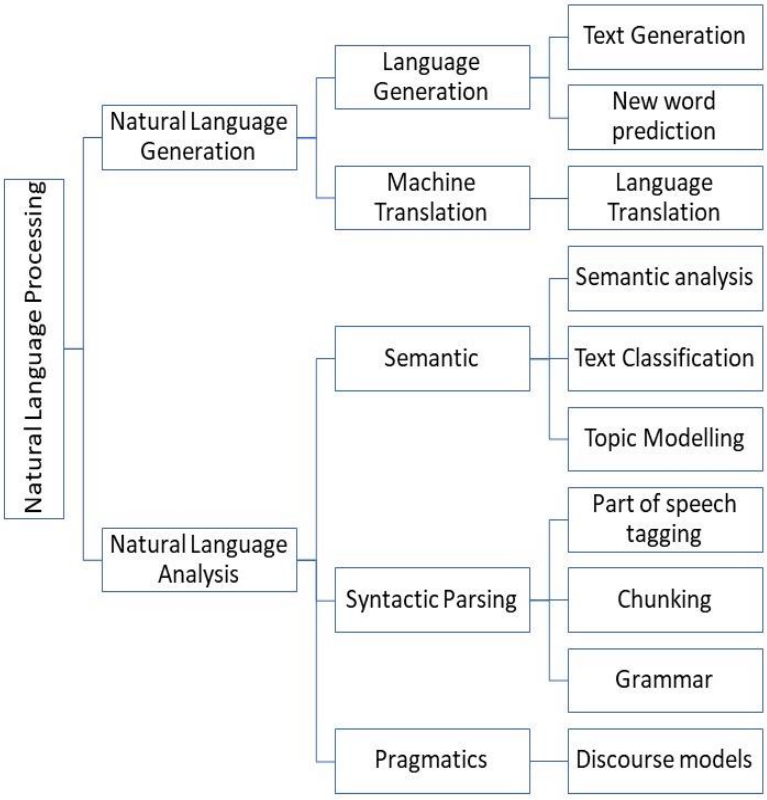

Figure 1 NLP Components

NLP allows human beings to interact with computers using a natural language just like human to human interaction. Thus, NLP combines different disciplines such as Artificial Intelligence, Human-Computer Interaction (HCI), Psychology, Philosophy, behavioral science and Mathematics.

NLP in Healthcare Systems

Natural Language Processing (NLP) has emerged as a transformative tool in modern healthcare, enabling the extraction of insights from the vast and growing volumes of unstructured clinical text. These texts include physician notes, pathology reports, and radiology findings to discharge summaries and mental health evaluations. Some of the applications of NLP in healthcare are shown in figure 2 and are briefly discussed as below:

Information Extraction

This is one of the most crucial and vital applications of NLP in healthcare domain to fetch information from text. This text is highly unstructured in nature and does not abide by grammatical rules and structures. It comprises of acronyms and abbreviations along with spelling and grammatical mistakes. NLP in this regard takes these clinical reports, nursing notes and user generated content and identifies the corresponding domains in the text and understands the linguistics and semantics of the sentences[10][11].

Clinical Information Retrieval

Information retrieval is the task of fetching information by searching. Clinical information retrieval is used by healthcare professionals for retrieving information about disease prevention, diagnosis and its treatment [12]. According to authors in CIRT is used for improving access to medical data and researches and provides improved decision making [13].

Medical document classification

Medical document classification makes use of NLP techniques to classify the document into predefined categories or topics. It includes topics such as Respiratory, Cardiology, digestive and Nephrology [14].

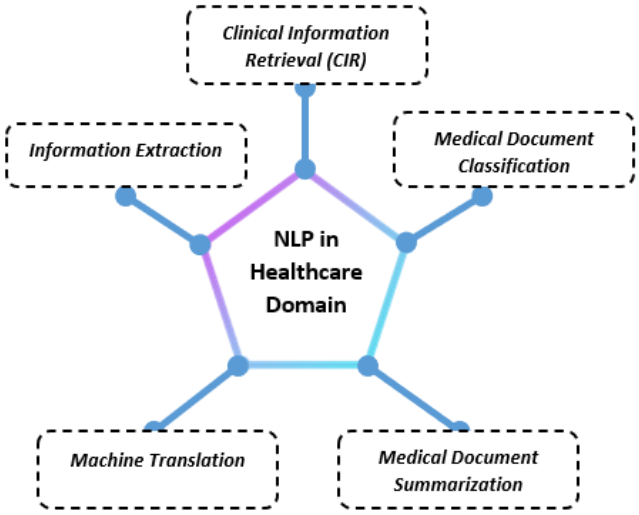

Figure 2: Applications of NLP in Healthcare Domain

Machine Translation

Machine translation is a subfield of artificial intelligence, it automatically converts one natural language to another. It allows large amount of data to be translated within a short span of time. It allows medical data translation of content such as patients complaints, adverse event etc. It's also used in clinical trials through neural machine translation. It produces fluent output language text [15].

Medical document summarization

Text summarization creates a summary of text from inputs such as clinical documents, discharge summary and nursing reports to make clinical decisions.

Case Study

NLP has many applications in the healthcare sectors. One of the applications being the efficacy of NLP in early detection of mental disorders such as depression in patients using clinical notes.

This section outlines the step by step approach which can be adopted for detection of early signs of depression. For illustrative purposes a small text has been used as data however, the same approach can be easily extended to real life and diverse clinical datasets.

In this case study we will be following a proposed systematic process shown in figure 3 to train machine learning and Deep Learning algorithms on the clinical notes.

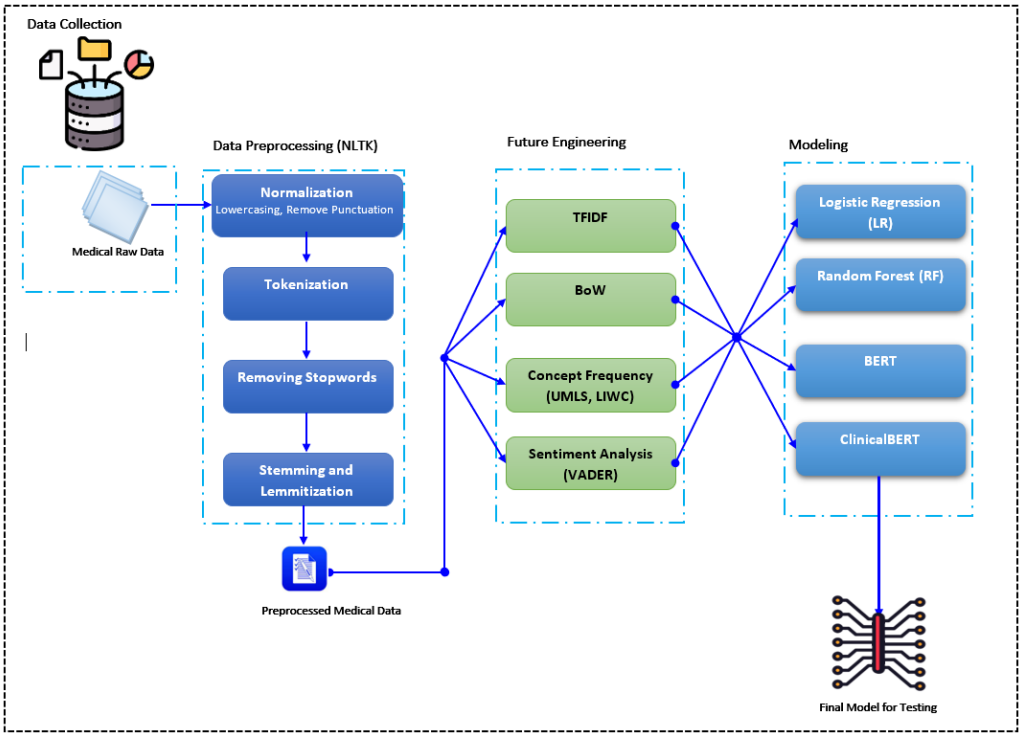

Figure 3: Proposed Systematic procedure for training a Model on Medical-Text Data

Data Collection

Effective data collection strategies need to be adopted to acquire data relating to detection of depression. Any data set comprising of clinical notes can be used. For example if we use a clinical note with the following text

Patient has Depression! Suicidal Ideation noted.

Patient presents with persistent sadness and fatigue.

The patient was experiencing sleeping difficulties.

No signs of mood disorder or depression reported.

Complains of fatigue and sleep disturbance for two weeks.

Regular check-up, no psychiatric concerns mentioned.

Diagnosed with major depressive disorder and started medication.

Preprocessing Techniques

Preprocessing is a critical step in any Natural Language Processing (NLP) application. healthcare data comprising of clinical notes are usually unstructured, noisy, and comprise of abbreviations, and spelling inconsistencies. Effective preprocessing methods helps standardize the data for better feature extraction and model performance. It therefore plays a vital role to transform samples of raw data into information. Preprocessing can include multiple steps such as

A. Normalization: The process of converting raw text into a consistent format.

- Lowercasing : Converts all text to lowercase to reduce dimensionality
- Remove Punctuation and Special Characters: Removes any punctuation marks or special characters which do not carry any semantic meaning for e.g., "!", "?", "#" can be removed [17].

For example:

Patient has Depression! Suicidal Ideation noted.

After normalization:

patient has depression suicidal ideation noted

B. Tokenization: It is the process of breaking a text string into individual units called tokens. It enables the identification of important words for analysis [17].

For example:

Patient presents with persistent sadness and fatigue.

After tokenization:

['Patient', 'presents', 'with', 'persistent', 'sadness', 'and', 'fatigue', '.']

C. Removal of stopwords: Stop words are common words such as "is," "the," "and". Since these words carry little meaningful information therefore, they are removed. Thus, this step reduces noise and computational complexity in text [18].

For example:

['Patient', 'presents', 'with', 'persistent', 'sadness', 'and', 'fatigue']

After stopword removal:

['patient', 'presents', 'persistent', 'sadness', 'fatigue']

***D.* Lemmatization:** It is the process of reducing words to their base or dictionary form or lemma. It improves generalization and helps models learn effectively.

For example:

The patient was experiencing sleeping difficulties.

After Lemmatization:

['the', 'patient', 'be', 'experience', 'sleep', 'difficulty', '.']

Feature Engineering

It is a crucial step in transforming raw and unstructured textual data into a numerical format that machine learning models can process effectively. For this case study, we used multiple feature engineering techniques that capture lexical, semantic, clinical, and affective information from the raw text. These techniques include :

Bag-of-Words (BoW)

The Bag-of-Words model serves as a foundational approach where each document is represented as a vector of word occurrence counts based on a fixed vocabulary. This model captures word presence and frequency but disregards grammar, order, and semantics. BoW provides strong baseline results, especially in traditional machine learning models like logistic regression and support vector machines[18]. For depression detection, It is effective for capturing direct mentions of depressive symptoms such as "depression", "fatigue", and "suicidal".

Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF builds on the BoW model by assigning importance to words based on their frequency within a specific document relative to their occurrence across the entire corpus. This technique helps highlight terms that are particularly meaningful for a given individual while reducing the influence of generic terms. In the context of depression detection, TF-IDF can help in reducing the weight of common terms like “patient” and emphasize n impactful terms like “anhedonia,” “hopelessness,” or “self-harm” which may serve as strong indicators of depressive states.

Concept Frequency using Lexical Resources:

To add interpretability and domain relevance, lexical features are extracted based on existing ontologies or psycholinguistic lexicons [19]. For example:

- UMLS (Unified Medical Language System) [19]: In clinical texts, depressive symptoms like “fatigue,” “insomnia,” and “hopelessness” can be mapped to Concept Unique Identifiers (CUIs), allowing for normalization and feature standardization.

- LIWC (Linguistic Inquiry and Word Count): LIWC maps words to psychological categories (e.g., "negative emotion," "anxiety," "cognitive processes") that are often associated with mental health conditions.

These concept frequencies are used as normalized or binary features to capture psychological or clinical relevance across texts.

Sentiment Analysis

Sentiment polarity and emotional tone are strong indicators of depression. Tools such as VADER (Valence Aware Dictionary for Sentiment Reasoning)[20] or TextBlob can be used to extract sentiment scores that reflect the overall affective state conveyed in a text.

Texts with consistently negative sentiment or polarity scores close to -1 may be indicative of depression.

Semantic Embeddings

Modern NLP methods use vector representations to capture the semantic and contextual meaning of words or entire documents. Embeddings allow the model to distinguish subtle contextual differences between clinical terms like “fatigue” and “tiredness” or “no psychiatric concerns” versus “suicidal ideation”. In this case study, the following embeddings were used:

- Word2Vec: Word2Vec learns word associations and can represent each document as the average of its word vectors. This enables capturing relationships like “sad” “unhappy” or “exhausted” or “tired.”

- BERT-based models: Contextual embeddings from pre-trained transformers such as BERT (general-purpose), BioBERT (biomedical domain) [21], and ClinicalBERT (a pretrained model for clinical text) [22] are employed to model the meaning of words based on their surrounding text.

Thus, such embeddings allow models to detect subtle linguistic cues associated with depressive language patterns, including contextual negations and co-occurrence of multiple symptoms.

Modelling and Performance Evaluations

To automatically classify textual input as indicative or non-indicative of depression, we experimented with a series of machine learning and deep learning models. The modeling phase involves training machine learning and deep learning algorithms on engineered features to detect signs of depression from clinical narratives. We have used here algorithms like Logistic Regression, Random forest classification, BERT and clinical BERT. The objective is to classify each clinical note as either “Depressive” or “Non-Depressive” and to identify the model that best captures depression-related language patterns across various contexts.

These models were trained on features derived through Bag-of-Words (BoW), TF-IDF, sentiment scores, and semantic embeddings such as Word2Vec and BERT-based models. Furthermore, all the models were evaluated using standard classification metrics such as accuracy, precision, recall, F1-Score and AUC-ROC.

Logistic Regression

Logistic Regression was chosen for its interpretability and efficiency on sparse data. It is a linear model that maps features to probabilities using a sigmoid function. Suitable for linearly separable problems. On performing Logistic regression following are the results obtained

Table 1: Metrices used for logistic regression

Metric	Value
Precision	0.67
Recall	1.00
F1-Score	0.80
ROC-AUC	1.00

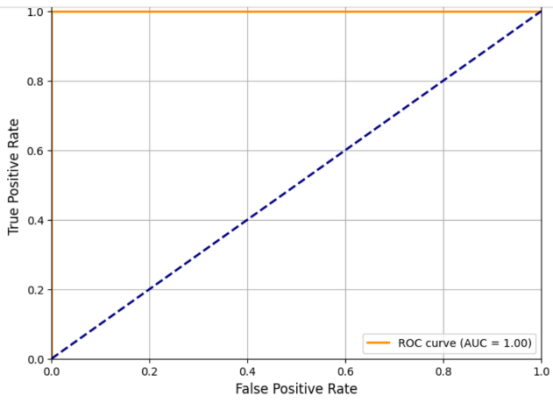

Figure 4: ROC Curve logistic regression

The logistic regression model was evaluated for its effectiveness in detecting depression-related symptoms from the example unstructured clinical note as shown in table 1. The model achieved a precision of 0.67, indicating that 67% of the clinical notes classified as indicative of depression were indeed associated with depression upon manual review. While this shows a moderate rate of false positives, it also suggests the model is reasonably accurate in its positive predictions.

The model attained a recall of 1.00, correctly identifying all instances that contained clinical indications of depression, such as "suicidal ideation," "persistent sadness," "fatigue," or a formal diagnosis (e.g., "major depressive disorder"). This high recall is especially critical in medical contexts, where failing to detect depressive symptoms could result in serious clinical oversight. The model’s ability to identify all positive cases suggests it is highly sensitive to language cues in clinical text.

The F1-score of 0.80 indicates a solid balance between precision and recall, reflecting the model's robustness in handling the trade-off between missing true cases and including false positives. Furthermore, the ROC-AUC of 1.00 suggests perfect discrimination between depression-positive and depression-negative cases in the test set. This exceptional score, while encouraging, should be interpreted with caution, as it may be indicative of potential overfitting since we have used very small data.

The ROC curve shown in figure 2 for the logistic regression model demonstrated a perfect classification performance, with an AUC (Area Under the Curve) of 1.00. This indicates that the model was able to flawlessly distinguish between clinical notes with and without depressive indicators. The curve reaches the top-left corner of the plot (TPR = 1, FPR = 0), reflecting maximum sensitivity and specificity.

While this result suggests excellent discriminatory ability, the unusually high AUC may also point to potential overfitting or dataset artifacts. This is due to the small sample size chosen.

However, the same can be applied to large dataset. Nonetheless, the model shows strong potential for automated detection of depression from clinical narratives

While this result suggests excellent discriminatory ability, the unusually high AUC may also point to potential overfitting or dataset artifacts. This is due to the small sample size chosen. However, the same can be applied to large dataset. Nonetheless, the model shows strong potential for automated detection of depression from clinical narratives.

Overall, the logistic regression model shows promise as a tool for automated identification of depressive symptoms from clinical narratives, with strong recall and acceptable precision. This can aid clinicians by flagging potentially high-risk patients based on free-text documentation, especially in resource-constrained settings.

Random Forest Classification

Random forest classification is an ensemble model combining multiple decision trees. It offers interpretability through feature importance. This algorithm when applied to the sample text generated the results as shown in table 2. The model achieved a Precision of 0.50, indicating that only 50% of the instances predicted as depressive were actually correct. This reflects a high rate of false positives. However, the Recall was 1.00, suggesting that the model successfully identified all actual depressive instances without missing any. While this high recall is desirable in clinical settings to ensure no cases are overlooked, it comes at the cost of precision.

The F1-Score, harmonizes precision and recall, was 0.67. This moderate value reveals an imbalance between correctly identifying depressive patients and minimizing false alarms. Most notably, the ROC-AUC score was 0.50, signifying that the model's ability to distinguish between depressive and non-depressive classes is equivalent to random guessing. This poor discriminative power is visually confirmed in the ROC curve shown in figure 3, which follows a diagonal line from the bottom-left to the top-right corner an indication of no effective classification boundary.

These results suggest that while the model is sensitive to detecting depressive cases, it lacks the specificity required for reliable prediction. The high recall but low precision and ROC-AUC highlight the model's tendency to overpredict depressive cases. This limitation could be due to the small and homogeneous dataset used. To enhance performance, more diverse and balanced clinical datasets can be used.

Table 2: Metrices used for Random forest

Metric	Value
Precision	0.500
Recall	1.000
F1-Score	0.670
ROC-AUC	0.500

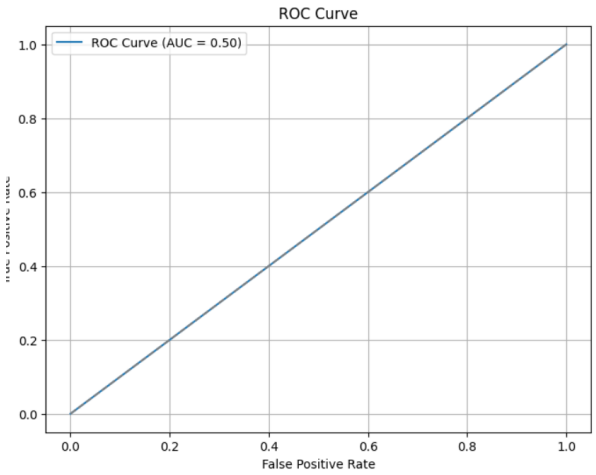

Figure 2 ROC Curve Random forest

LSTM (Long Short-Term Memory)

LSTM (Long Short-Term Memory) is a type of recurrent neural network designed to capture long-term dependencies in sequential data, making it well-suited for processing clinical text.

In our experiment, the LSTM model as shown in table 3 achieved a perfect recall (1.000) and ROC-AUC (1.000), meaning it correctly identified all positive cases and separated classes perfectly.

Table 3: Metrices used for LSTM model

Metric	Value
Precision	0.670
Recall	1.000
F1-Score	0.800
ROC-AUC	1.000

With a precision of 0.670 and an F1-score of 0.800, it shows strong overall performance, though the moderate precision suggests some false positives. These results indicate LSTM is highly effective for identifying depression symptoms in clinical notes but may benefit from fine-tuning to improve precision

ClinicalBERT / BERT

Transformer models like BERT capture rich, contextual semantics, particularly useful for subtle depressive cues. ClinicalBERT is a domain-specific version of BERT pre-trained on MIMIC-III and clinical notes [22].

Table 4: Metrices used for ClinicalBERT model

Metric	Value
Precision	1.000
Recall	0.400
F1-Score	0.571
ROC-AUC	0.800

It is a transformer-based model pre-trained on biomedical and clinical text, designed to understand complex medical language. In this study, ClinicalBERT as shown in table 4 achieved perfect precision (1.000), meaning all predicted positives were correct, but its recall was low (0.400), indicating it missed many actual positive cases. The F1-score of 0.571 and ROC-AUC of 0.800 reflect a conservative prediction approach. While the model is highly accurate when it does predict a positive case, its low recall suggests the need for better fine-tuning or threshold adjustment to improve its sensitivity in detecting clinical depression cues.

Results and Analysis

This section presents the outcomes of our experimental evaluation, focusing on the comparative performance of multiple Natural Language Processing (NLP) pipelines and classification models in detecting depressive symptoms from clinical text. It should be noted here that we are presenting the above mentioned results again in order to provide a comparative analysis within the selected machine and deep learning architectures.

Table 5: Performance Evaluation of Different Models on Clinical Notes

Model	ROC-AUC	Precision	Recall	F1-score
Logistic Regression	1.000	0.670	1.00	0.800
Random Forest	0.500	0.500	1.00	0.670
LSTM	1.000	0.670	1.000	0.800
ClinicalBERT	0.800	1.000	0.400	0.571

As presented in the table 5 the results from the classification of clinical notes using different models reveal interesting contrasts in model behavior and effectiveness. Logistic Regression and LSTM both achieved perfect ROC-AUC scores of 1.000, with identical precision (0.670), recall (1.000), and F1-score (0.800). These numbers suggest that both models are capable of identifying all relevant positive cases (perfect recall), while maintaining a moderate precision. However, the perfection in ROC-AUC is unusual and could indicate overfitting or data leakage, especially given that such performance is rarely observed in real-world clinical text classification tasks unless the dataset is small or the features are overly optimized.

Random Forest, on the other hand, performs poorly with a ROC-AUC of 0.500, which is no better than random guessing. While its recall is also 1.000, the precision is low at 0.500, suggesting that it predicts nearly everything as positive. This behavior results in a relatively lower F1-score of 0.670 and highlights a lack of discriminative power, possibly due to the high dimensionality or sparsity of the clinical text features that Random Forests typically struggle with.

ClinicalBERT presents a very different profile. Its ROC-AUC of 0.800 is respectable and indicates a good ability to separate the classes. It achieves perfect precision, meaning every positive prediction it makes is correct. However, its recall is only 0.400, meaning it misses a significant number of actual positives. This leads to a relatively low F1-score of 0.571. ClinicalBERT's conservative nature only labeling as positive when highly confident—could be due to insufficient fine-tuning, a poorly chosen classification threshold, or an imbalanced dataset where the model learns to be risk-averse in predicting positives.

Overall, while Logistic Regression and LSTM appear to perform exceptionally well, their results may not generalize due to potential overfitting.

Table 6: Classification accuracy of selected models on Clinical Samples.

Model	Accuracy
Logistic Regression	0.674%
Random Forest	0.50%
LSTM	0.67%
ClinicalBERT	0.571%

ClinicalBERT, though more cautious, demonstrates potential with its high precision and solid ROC-AUC, suggesting that with additional fine-tuning and threshold adjustment, it could strike a better balance between precision and recall. These findings highlight the importance of interpreting model metrics holistically, particularly in sensitive domains like medical data where false negatives and false positives carry significant implications. Similarly, we have presented accuracy scores for the classification of clinical notes in the table 6. The table show that Logistic Regression and LSTM performed similarly well, both achieving around 67%. This indicates that these models were able to correctly classify the majority of the instances and suggests their suitability for handling structured text features or sequence-based patterns in medical data. Their relatively strong performance, despite being simpler compared to deep contextual models, highlights the effectiveness of well-tuned traditional and sequential neural models on certain clinical datasets.

In contrast, Random Forest achieved only 50% accuracy, reflecting random performance and indicating that it struggled with the structure and complexity of clinical text. ClinicalBERT, though designed specifically for biomedical language understanding, achieved a slightly better accuracy of 57%, which is lower than both Logistic Regression and LSTM. This may be due to limited fine-tuning or a mismatch between the model's complexity and the size or nature of the dataset. While ClinicalBERT has potential, it likely needs further optimization to outperform simpler models in this context.

Conclusion and Future Directions

This paper explored the application of Natural Language Processing (NLP) in the healthcare domain, focusing on the classification of clinically significant information embedded within unstructured clinical texts such as physician notes and discharge summaries. By implementing a practical case study aimed at detecting early signs of depression from real-world clinical notes, we assessed the comparative effectiveness of classical machine learning models (Logistic Regression, Random Forest), deep learning methods (LSTM), and transformer-based approaches (ClinicalBERT). Our findings revealed that Logistic Regression and LSTM performed similarly and achieved higher accuracy than Random Forest and even ClinicalBERT in this specific task. While ClinicalBERT demonstrated high precision, its lower recall and accuracy highlighted the need for more task-specific fine-tuning and threshold optimization to unlock its full potential.

These outcomes indicate that simpler models can still be highly effective in certain clinical NLP tasks, especially when data availability or computational resources are limited. Future research should focus on optimizing transformer models like ClinicalBERT for domain-specific classification tasks through improved training strategies, hyperparameter tuning, and threshold calibration. Further work is also needed to evaluate model robustness across different datasets and clinical settings. Integrating multimodal data—combining unstructured text with structured inputs like lab results and imaging—could enhance the diagnostic context and model accuracy. Additionally, embedding these NLP tools into clinical decision support systems (CDSS), along with explainable AI techniques, would promote transparency and facilitate real-world adoption. Ensuring patient privacy through

federated learning and addressing algorithmic fairness will also be crucial for building equitable and trustworthy NLP solutions in healthcare.

References

- J. Friedman, R. G. Parrish, and D. A. Ross, "Electronic Health Records and US Public Health: Current Realities and Future Promise," Am J Public Health, vol. 103, no. 9, p. 1560, Sep. 2013, doi: 10.2105/AJPH.2013.301220.
- N. Afzal et al., "Natural language processing of clinical notes for identification of critical limb ischemia," Int J Med Inform, vol. 111, pp. 83–89, Mar. 2018, doi: 10.1016/j.ijmedinf.2017.12.024.
- K. Savova et al., "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications," J Am Med Inform Assoc, vol. 17, no. 5, p. 507, Sep. 2010, doi: 10.1136/JAMIA.2009.001560.
- J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, vol. 1, pp. 4171–4186, Oct. 2018, Accessed: May 20, 2025. [Online]. Available: https://arxiv.org/pdf/1810.04805
- W. W. Chapman, P. M. Nadkarni, L. Hirschman, L. W. D'Avolio, G. K. Savova, and O. Uzuner, "Overcoming barriers to NLP for clinical text: The role of shared tasks and the need for additional creative solutions," Journal of the American Medical Informatics Association, vol. 18, no. 5, pp. 540–543, Sep. 2011, doi: 10.1136/AMIAJNL-2011-000465,.
- Y. Wang et al., "Clinical information extraction applications: A literature review," J Biomed Inform, vol. 77, pp. 34–49, Jan. 2018, doi: 10.1016/j.jbi.2017.11.011.
- E. Hossain et al., "Natural Language Processing in Electronic Health Records in relation to healthcare decision-making: A systematic review," Comput Biol Med, vol. 155, Mar. 2023, doi: 10.1016/j.compbiomed.2023.106649.
- S. Wu et al., "Deep learning in clinical natural language processing: A methodical review," Journal of the American Medical Informatics Association, vol. 27, no. 3, pp. 457–470, Mar. 2020, doi: 10.1093/JAMIA/OCZ200,.
- Le Glaz et al., "Machine Learning and Natural Language Processing in Mental Health: Systematic Review," J Med Internet Res 2021;23(5):e15708 https://www.jmir.org/2021/5/e15708, vol. 23, no. 5, p. e15708, May 2021, doi: 10.2196/15708.
- S. Perera, A. Sheth, K. Thirunarayan, S. Nair, and N. Shah, "Challenges in understanding clinical notes: Why NLP engines fall short and where background knowledge can help," International Conference on Information and Knowledge Management, Proceedings, pp. 21–26, 2013, doi: 10.1145/2512410.2512427.
- Elbattah, M., Arnaud, É., Gignon, M., & Dequen, "The Role of Text Analytics in Healthcare: A Review of Recent Developments and Applications," HEALTHINF, pp. 825–832, 2021.
- M. P. Gagnon et al., "A systematic review of interventions promoting clinical information retrieval technology (CIRT) adoption by healthcare professionals," International Journal of Medical Informatics, vol. 79, no. 10, pp. 669–680, Oct. 2010, doi: 10.1016/J.IJMEDINF.2010.07.004.
- P. Pluye, R. M. Grad, L. G. Dunikowski, and R. Stephenson, "Impact of clinical information-retrieval technology on physicians: a literature review of quantitative, qualitative and mixed methods studies," International journal of medical informatics, vol. 74, no. 9, pp. 745–768, 2005, doi: 10.1016/J.IJMEDINF.2005.05.004.
- S. Kashyap, S. M. G, V. Rajaram, and V. S, "Medical Document Classification – IJERT," International Journal of Engineering Research & Technology (IJERT), pp. 1–4, 2015.
- Stanford, "The Stanford Natural Language Processing Group." Accessed: Apr. 23, 2022. [Online]. Available: https://nlp.stanford.edu/projects/mt.shtml
- S. Cohen, K. R. Mitchell, and B. Elvevåg, "What do we really know about blunted vocal affect and alogia? A meta-analysis of objective assessments.," Schizophrenia Research, vol. 159, no. 2–3, pp. 533–538, Sep. 2014, doi: 10.1016/J.SCHRES.2014.09.013.
- S. Bird, E. Klein, and E. Loper, "LIVRO: cookbook Natural Language Processing with Python," J Endod, vol. 28, no. 4, pp. 330–332, 2009, Accessed: May 30, 2025. [Online]. Available: https://www.oreilly.com/library/view/natural-language-processing/9780596803346/
- D. Manning, P. Raghavan, and H. Schütze, "Introduction to Information Retrieval," Introduction to Information Retrieval, Jul. 2008, doi: 10.1017/CBO9780511809071.

- R. Aronson and F. M. Lang, “An overview of MetaMap: Historical perspective and recent advances,” Journal of the American Medical Informatics Association, vol. 17, no. 3, pp. 229–236, May 2010, doi: 10.1136/JAMIA.2009.002733,.
- J. Hutto and E. Gilbert, “VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text,” Proceedings of the International AAAI Conference on Web and Social Media, vol. 8, no. 1, pp. 216–225, May 2014, doi: 10.1609/ICWSM.V8I1.14550.
- J. Lee et al., “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” Bioinformatics, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, doi: 10.1093/BIOINFORMATICS/BTZ682.
- Alsentzer et al., “Publicly Available Clinical BERT Embeddings,” pp. 72–78, Jul. 2019, doi: 10.18653/V1/W19-1909.