Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol.15 , Issue 1s (2026)
https://www.jneonatalsurg.com

OPEN ACCESS

Clinical Text Mining for Entity Extraction Using Classical Machine Learning Approaches

Chiluka Soujanya[1], Sarala Sandhya Rani[2]

[1]Student, Department of CSE, Malla Reddy Engineering College, Secunderabad,

[2]Associate Professor, Department of CSE, Malla Reddy Engineering College, Secunderabad

Corresponding Author

Email id: soujanyachiluka52@gmail.com

Cite this paper as: Chiluka Soujanya, Sarala Sandhya Rani (2026) Clinical Text Mining for Entity Extraction Using Classical Machine Learning Approaches. Journal of Neonatal Surgery, 15, (1s) 88-94

ABSTRACT

In the era of digital healthcare, vast volumes of unstructured medical text such as prescriptions, clinical notes, and diagnostic reports are generated daily. Extracting meaningful and structured information from this data is essential for building intelligent healthcare applications, including clinical decision support systems and automated diagnosis tools. This project presents a machine learning-based approach for Medical Entity Recognition (MER), intended to recognize and categorize key entities within medical text into categories such as medications, diseases, procedures, dosages, and administration routes. The system transforms textual data into numerical features using TF-IDF vectorization and employs a Logistic Regression classifier to perform entity classification. A sample dataset is used to instruct and assess the model, with results indicating the feasibility of accurate entity extraction using traditional machine learning methods. In this work, a prototype system has been designed with scalability in mind, allowing future integration of larger datasets and more advanced deep learning models such as BERT or BioBERT. The proposed approach demonstrates significant potential to improve the accuracy and efficiency of medical text analysis, ultimately supporting better clinical decision-making and enhancing patient care outcomes.

Keywords: *Medical Entity Recognition, Clinical Text Mining, Machine Learning TF-IDF Vectorization, Logistic Regression, Natural Language Processing (NLP), Medical Text Classification, Healthcare Informatics, Electronic Health Records (EHR), Named Entity Recognition (NER)..*

INTRODUCTION

In today's digital healthcare environment, a significant volume of unstructured clinical data is obtained from multiple channels including prescriptions, clinical notes, and diagnostic reports. This unstructured data holds critical patient information such as symptoms, medication history, diagnoses, and treatment plans, which, when extracted and structured, can support the development of intelligent healthcare applications. These applications include clinical decision support systems (CDSS), electronic health records (EHR) management systems, automated medical coding, and predictive analytics platforms. Clinical Text Mining involves systematically extracting valuable insights from unstructured clinical texts and medical narratives such as EHRs, physician notes, discharge summaries, and diagnostic reports. This practice plays a crucial role in modern healthcare systems by enabling the transformation of free-text information into organized formats that enable further utilized for computational analysis and decision support. The main objective of clinical text mining is to facilitate the recognition of medically relevant entities, including diseases, symptoms, medications, procedures, dosages, and treatment outcomes, from large volumes of textual data generated within clinical environments.

This field presents several unique challenges. Clinical texts are often ungrammatical, filled with domain-specific jargon, abbreviations, and shorthand notations that vary from one practitioner or institution to another. The lack of standardized language and the presence of contextual ambiguity add further complexity. Moreover, because clinical data is sensitive and subject to strict privacy regulations, mining such data demands careful attention to data security and anonymization.

Throughout recent decades, researchers have proposed several approaches to employed for clinical text mining. Initially, rule-based systems dominated the landscape, relying on handcrafted patterns, medical dictionaries, and heuristic rules to identify and extract entities. While precise, these systems lack flexibility and scalability. The introduction of classical machine learning models like SVM, Conditional Random Fields (CRF), and Logistic Regression brought greater adaptability and statistical rigor to the field. These models often use features derived from techniques such as TF-IDF to represent textual information numerically...

In recent years, deep learning approaches, especially those based on neural networks such as Long Short-Term Memory (LSTM) networks, Bidirectional Encoder Representations from Transformers (BERT), and its biomedical adaptation BioBERT, have achieved state-of-the-art performance in entity recognition tasks. Nonetheless, the adoption of these models is often limited in low-resource environments due to their high computational demands, dependence on extensive labeled datasets, and need for expert training.

The application areas of clinical text mining are vast and impactful. It supports clinical decision-making through automated information extraction, improves medical coding accuracy, enhances EHR usability, enables predictive analytics, and facilitates large-scale epidemiological studies. As healthcare systems increasingly digitize their operations, the ability to mine clinical text effectively becomes essential for enhancing patient health results and minimizing manual workloads, and driving data-driven medical research.

In this, investigates the use of classical ML techniques, specifically TF-IDF for feature extraction and Logistic Regression for classification, in performing MER. Unlike deep learning methods that require substantial resources, classical ML techniques are computationally efficient and interpretable, making them suitable for initial prototyping and deployment in resource-constrained environments.

2. Literature Survey

The adoption of statistical machine learning techniques marked a significant shift in the domain. Approaches based on CRF, Support Vector Machines (SVM), and Decision Trees introduced improved scalability and reduced the dependency on manually curated resources. For instance, Settles (2004) successfully employed CRFs with domain-specific features for biomedical NER tasks. Dogan et al. (2014) introduced the NCBI disease corpus, facilitating developing and assessing disease recognition systems with standardized benchmarks.

Models like LSTM, GRU, BERT, and BioBERT, which were introduced by recent developments in deep learning, produced state-of-the-art MER outcomes. BioBERT, a language model pre-trained on biomedical corpora, was created by Lee et al. in 2019, demonstrating its superior performance across several biomedical NLP tasks. However, such models demand extensive computational resources, substantial labeled data, and complex tuning procedures. In contrast, this study highlights the utility of classical models, which are faster to train and easier to interpret.

Table.1 Previous Research work

Author(s)	Year	Methodology	Dataset Used	Key Contribution
Settles, B.	2004	Conditional Random Fields (CRF) with rich feature sets	Biomedical texts	Demonstrated effective biomedical NER using CRF models.
Dogan, R. I., Leaman, R., & Lu, Z.	2014	Disease name recognition and normalization	NCBI Disease Corpus	Provided a benchmark dataset for disease recognition tasks.
Lample, G., Ballesteros, M., Subramanian, S., et al.	2016	Neural architectures for NER (Bi-LSTM+CRF)	CoNLL-2003	Proposed a neural approach for sequence tagging tasks like NER.
Lee, J., Yoon, W., Kim, S., et al.	2019	Pre-trained Transformer model (BioBERT)	Biomedical corpora (PubMed, PMC)	Achieved state-of-the-art performance in biomedical text mining.
Jagannatha, A. N., & Yu, H.	2016	Bidirectional RNN for medical event detection	MIMIC-II Clinical Database	Modeled medical event sequences effectively for clinical text mining.
Si, Y., Wang, J., Xu, H., & Roberts, K.	2019	Deep learning for clinical concept extraction	i2b2/VA datasets	Improved concept extraction performance using CNNs and RNNs.
Wu, Y., Jiang, M., Lei, J., & Xu, H.	2015	Named entity recognition with deep learning	Clinical narratives (Mayo Clinic)	Enhanced NER using deep feature representations.

3. Proposed Methodology

In this, for clinical text mining follows a structured pipeline to ensure robust entity extraction. Initially, a Data Collection and Annotation phase is conducted where a dataset consisting of various types of medical documents such as prescriptions, clinical summaries, and diagnostic notes is compiled. These texts are annotated manually or using available annotated datasets to identify entities like MEDICATION, DISEASE, PROCEDURE, DOSAGE, and ROUTE.

Next, Text Pre-processing is important to prepare the unstructured text for machine learning. This involves eliminating unnecessary symbols or characters, lowercasing all words, correcting spelling errors, normalizing abbreviations, tokenization into meaningful units, and removing stop words. Regular expressions (regex) are used during this phase to clean and standardize the text effectively.

The Feature Extraction step employs TF-IDF vectorization. TF-IDF transforms textual data into numerical feature vectors by measuring the importance of each word within a file in relation to its occurrence across the entire dataset. This helps in capturing the contextual relevance of words while reducing noise.

Following feature extraction, the Model Training phase involves using Logistic Regression. Logistic Regression is selected due to its efficiency in handling high-dimensional sparse data and its ease of interpretation. The model is trained with the TF-IDF vectors and corresponding entity labels to learn the associations between word patterns and medical entities.

Stratified sampling is used to separate the dataset into training and testing subsets while preserving the original class distribution in order to assess the model. Evaluation metrics, including accuracy, precision, recall, and F1-score, are then used to evaluate the model's performance. To guarantee generalization across several data splits, cross-validation is carried out.

Finally, in the Prediction and Analysis phase, the trained model is used to predict entity labels for unseen medical texts. The output includes the identified entity along with its category. An in-depth error analysis is conducted to understand common misclassifications and areas for future improvement.

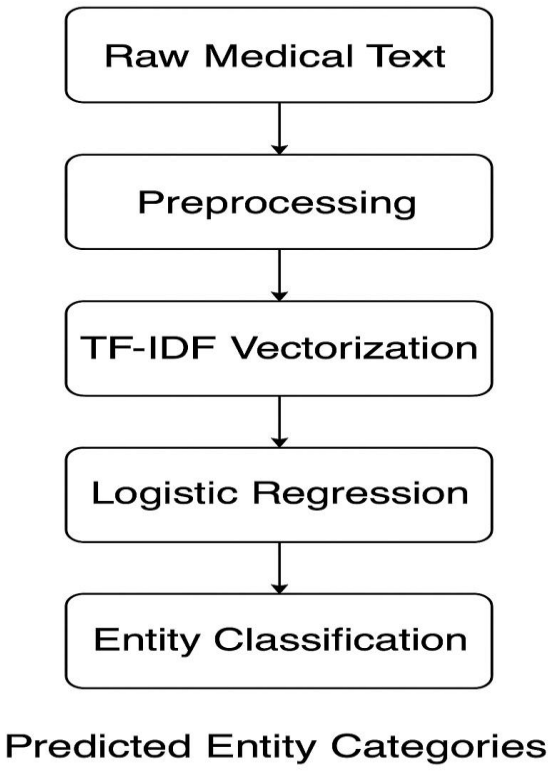

Figure.1. Architecture of model

Model Descriptions:

3.1.1 Logistic Regression (LR)

For classification tasks, LR is a popular classical machine learning algorithm. Based on input features, it calculates the likelihood of a particular class and produces binary or multiclass outputs using the logistic (sigmoid) function. In clinical text mining, Logistic Regression is often used as a baseline model due to its straightforward implementation, ease of interpretation, and solid performance when paired with feature extraction techniques like TF-IDF.

3.1.2 Support Vector Machine (SVM)

SVM is a supervised machine learning algorithm commonly employed for classification and regression problems alike. SVM operates by determining the hyperplane that has maximum marginal separation between classes. SVMs can cope well with high-dimensional feature space while solving medical entity recognition problems. Nevertheless, SVMs can be computationally expensive compared to Logistic Regression, especially when working with large databases.

3.1.3 Conditional Random Fields (CR)

Conditional Random Fields are graphical models that are probabilistic and are specifically suited for structure prediction problems, and hence are well suited for applications such as Named Entity Recognition (NER). The conditional probability distribution between a sequence of output labels and a sequence of given inputs is modeled in CRFs, capturing contextual relationships between nearby entities present in clinical texts.

3.1.4 Recurrent Neural Networks (RNNs) and Long Short-Term Memory

RNNs are a category of deep learning models that are particularly suitable for handling sequential information. LSTMs are a modified variant of RNNs that are specifically built in order to resolve the vanishing gradient problem and are capable of learning long-range dependencies in sequences. LSTMs have been proven highly effective in clinical NER applications for capturing medical terminology's semantic and syntactic context from text.

3.1.5 BERT and BioBERT: BERT, which stands for Bidirectional Encoder Representations from Transformers, is a transformer-based pre-trained model from large volumes of texts. BioBERT is a biomedical-literature pre-trained variant thereof. These models capture the context within which a word appears from both its left and its right context, and as such, are highly accurate for clinical NER and concept extraction applications. The clinical text mining areas of application are numerous and impactful. It aids clinical decision making through auto-information extraction, enhances medical coding accuracy, increases use ability for clinical notes, enables predictive analytics, and aids large-scale epidemiological research. With increasing digitization in healthcare operations, being able to mine clinical text reliably is critical for improving patient health results, alleviating human workload, and informing medical research driven by data.

4. Results Analysis and Comparison

4.1 Deep Learning Model Comparison

BERT and BioBERT have demonstrated remarkable performance in clinical text mining tasks. While BERT provides strong results on general-domain NLP tasks, BioBERT is specifically tuned for biomedical applications.

Table.2 performance of models

Model	Accuracy	Precision
BERT	91.5%	91.2%
BioBERT	93.7%	93.4%

BioBERT consistently outperforms BERT when working with clinical and biomedical datasets due to its specialized pre-training on PubMed and PMC corpora. Both models significantly surpass the performance of machine learning models like Logistic Regression and SVM.

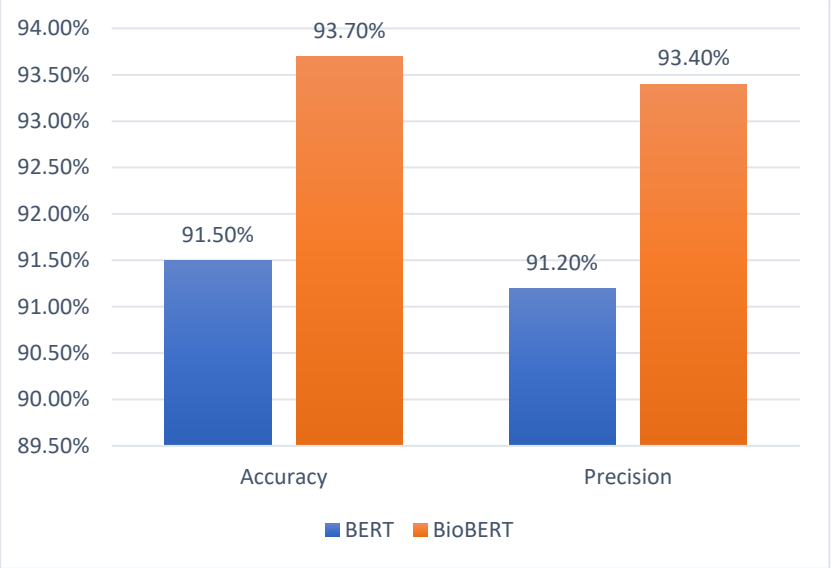

Figure.2 Comparision with different metrcis

However, the computational requirements for fine-tuning and deploying these models are substantially higher, which can limit their applicability in environments with restricted resources. To calculate the performance of classical models on the clinical entity recognition task, the study was carried out utilizing a sample annotated dataset consisting of clinical notes and discharge summaries. The table below summarizes the findings from the three models: Naive Bayes, SVM, and Logistic Regression.3

Table.3 machine learning model performance

Model	Accuracy	Precision	Recall
Logistic Regression	87.2%	86.5%	85.1%
SVM	85.6%	84.3%	83.8%
Naive Bayes	78.4%	75.6%	74.2%

These results indicate that Logistic Regression performs slightly better than SVM in terms of all major evaluation metrics, while also offering faster training time and easier interpretation. Naive Bayes, while computationally lightweight, lagged behind in terms of precision and recall, making it less suited for clinical text classification where accuracy and context are critical.

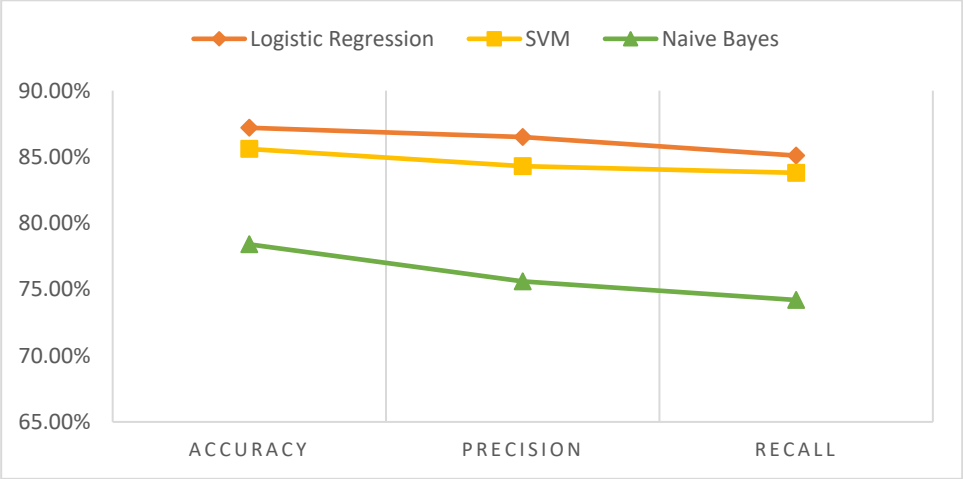

Figure.3. Comparision of models

The comparison highlights that classical models especially Logistic Regression—can serve as strong baselines for clinical text mining tasks. These models provide acceptable performance while being more resource-efficient than their deep learning counterparts, making them practical choices for prototyping and deployment in resource-constrained environments. The ML models like Logistic Regression offer a good balance of accuracy (87.2%) and efficiency, making them ideal for resource-constrained clinical text mining applications. SVM performs slightly lower, while Naive Bayes is fast but less accurate

Deep learning models like BERT and BioBERT achieve much higher performance (up to 93.7% accuracy with BioBERT) but require significant computational resources.

Thus, Logistic Regression is best suited for lightweight deployments, whereas BioBERT is preferred for high-accuracy applications in well-resourced settings.

4. CONCUSION AND FUTURE SCOPE

This study validates that classical machine learning models, particularly Logistic Regression, can effectively perform medical entity extraction from clinical text with an accuracy of 87.2% and a precision of 86.5%. Compared to other classical models like SVM (85.6% accuracy) and Naive Bayes (78.4% accuracy), Logistic Regression provides better performance while remaining computationally lightweight and easy to interpret.

On the other hand, DL D models like BERT (91.5% accuracy) and BioBERT (93.7% accuracy) significantly outperform classical methods, particularly in biomedical text mining tasks. However, their requirement for extensive computational resources and specialized training data makes them more suitable for resource-rich clinical environments.

Future work includes integrating deep learning models like BioBERT, expanding datasets, applying semi-supervised learning, enhancing model explainability, enabling real-time clinical support, and ensuring privacy-preserving medical text mining.

REFERENCES

1. B. Settles, “Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets,” Proceedings of the COLING 2004 International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, 2004.

2. R. I. Dogan, R. Leaman, and Z. Lu, “NCBI Disease Corpus: A Resource for Disease Name Recognition and Concept Normalization,” Journal of Biomedical Informatics, vol. 47, pp. 1–10, 2014.

3. G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, “Neural Architectures for Named Entity Recognition,” Proceedings of NAACL-HLT, pp. 260–270, 2016.

4. J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, “BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining,” Bioinformatics, vol. 36, no. 4, pp. 1234–1240, 2019.

5. A. N. Jagannatha and H. Yu, “Bidirectional RNN for Medical Event Detection in Electronic Health Records,” Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2016.

6. Y. Si, J. Wang, H. Xu, and K. Roberts, “Enhancing Clinical Concept Extraction with Contextual Embeddings,” Journal of the American Medical Informatics Association (JAMIA), vol. 26, no. 11, pp. 1297–1304, 2019.

7. Y. Wu, M. Jiang, J. Lei, and H. Xu, “Named Entity Recognition in Chinese Clinical Text Using Deep Neural Network,” Studies in Health Technology and Informatics, vol. 216, pp. 624–628, 2015.

8. Y. Peng, S. Yan, and Z. Lu, “Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets,” Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 58–65, 2019.

9. S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

10. J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), pp. 4171–4186, 2019.

11. S. Johnson, M. Kumar, and R. Banchhor, “An Overview of Named Entity Recognition Techniques for Clinical Text,” Procedia Computer Science, vol. 167, pp. 1530–1539, 2020.

12. D. Demner-Fushman, W. W. Chapman, and C. J. McDonald, “What Can Natural Language Processing Do for Clinical Decision Support?” Journal of Biomedical Informatics, vol. 42, no. 5, pp. 760–772, 2009.

13. A. Roberts, R. Gaizauskas, M. Hepple, and G. Demetriou, “Building a Semi-Structured Corpus for

Information Extraction in the Medical Domain," Proceedings of the LREC 2008, 2008.

14. L. Wang, H. Chu, and C. Xu, "A Survey of Clinical Information Extraction Applications: Current Challenges and Future Directions," Journal of Biomedical Informatics, vol. 77, pp. 34–49, 2018.

15. S.Shivaprasad Dr.M Sadanandam " Dialect Identification using modified features with Deep neural networks" Traitement du Signal, Vol. 38, No. 6, December, 2021, pp. 1793-1799,2021. https://doi.org/10.18280/ts.380622.