

A Biomedical NLP Pipeline Including Temporal Parsing, Syntactic Relation Mining and Domain-Specific Embeddings

Tasnim Fariha^{1*}, Jasiya Fairiz Raisa¹

Abstract

This study presents a comprehensive medical natural language processing pipeline that combines advanced entity extraction, syntactic relation mining, domain-specific embeddings and temporal parsing to transform annotated biomedical abstracts into structured clinical information. The pipeline framework uses a modular workflow that includes data parsing, spaCy-based preprocessing, exploratory corpus analysis, BioBERT-driven NER, dependency-based relation extraction and rule-based temporal detection to handle 593 annotated abstracts with inline XML-style biological entities. A BioBERT-extracted collection of 3,490 entity mentions from the first 100 documents is produced by the pipeline, which identifies 4,424 annotated entities with disease-related categories predominating the corpus structure. Mapping disease-gene-treatment relationships with a limited number of temporal indications is efficient, as demonstrated by the 613 syntactic relationships generated by relation extraction and the 28 duration expressions captured by temporal parsing. According to co-occurrence analysis, there is a substantial correlation between modifier and specific disease, which reflects typical clinical narrative patterns. This study presents a biomedical NLP pipeline for extracting entities, finding relations and parsing time. It aims to transform annotated biomedical text into structured clinical knowledge. However, it also points out limitations in time coverage and adaptability.

Keywords: NLP, Clinical Text, EHR, Context-aware, NCBI corpus

1. Introduction

The recent advancement in Electronic Health Records (EHRs) has led to extensive unstructured clinical text that consists of clinician notes, discharge summaries and diagnostic reports. These data types have EHRs that predict the

*Corresponding Author email: tasnimh991@gmail.com

¹Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh.

DOI: <https://doi.org/10.64494/JFST/v4i1/tf/2026/06/109-124>

future of healthcare through data analytics. Despite the significance of EHRs, the specialized and complex language of clinical narratives presents an almost impenetrable challenge to conventional natural language processing (NLP) techniques. This has highlighted the need for more advanced models that are especially designed for the biomedical domain (Lee et al., 2020). In comparison to traditional NLP models, some recent developments that are domain-specific and pre-trained language models (PLMs) such as BioBERT and ClinicalBERT, are better at understanding clinical language and reasoning. For example, in clinical text mining, BioBERT greatly improves the tasks of named entity recognition (NER) and relation extraction, owing to its ability to capture the intensive professional medical language and complex contextual relations after being pre-trained on large biomedical data. Moreover, studies demonstrate the importance of temporal reasoning, when combined with semantic role to predict the progress of diseases and the course of action to treat such diseases, in a temporal manner. Given specialized terminology, implicit relations, fragmented temporal signals and task-isolated extraction methodologies that do not maintain full clinical context, it is still challenging to extract integrated, context-rich clinical knowledge from unstructured biomedical material, even with advancements in natural language processing. Most existing research on biomedical natural language processing processes the three core tasks of named entity recognition, relation extraction and temporal analysis independently and can only recognize simple descriptions such as "the patient has diabetes". This study integrates these three tasks to build a unified pipeline, filling the research gap in this field. To address these limitations, this study focuses on the following three research challenges:

- a. General-purpose and domain-adapted NLP models often fail to capture full syntactic, semantic and temporal context in clinical text; for example, the phrase "post-treatment improvement over 3 months" may not be correctly linked to both treatment and its associated timeline, leading to incomplete disease progression modelling due to implicit relations and sparse temporal signals.
- b. Current biomedical NLP pipelines treat named entity recognition, relation extraction and temporal modelling as 3 independent tasks, thereby leading to fragmented clinical narratives. The inability to integrate information from disparate sources in an end-to-end manner results in inadequate representations of disease–gene–treatment interactions and hinders the interpretability and clinical utility of extracted knowledge.

- c. Annotated biomedical datasets like the NCBI Disease Corpus are often used to evaluate named entity recognition. However, they are seldom examined for extracting knowledge that is contextual, relational and at the corpus level.

The major novelties and contributions from this study are highlighted in the following detailed points:

- a. This work describes a linear, modular, domain-specific NLP pipeline that processes biomedical text. This approach follows the more general trend of using transformer-based architectures, which have continuously shown better contextual representation in a variety of NLP applications (Mirza et al., 2025). It includes parsing, preprocessing, named entity recognition, relation selection and temporal expression parsing within one framework for extracting structured clinical knowledge.
- b. BioBERT-based embeddings improve the ability to identify disease-related entities. Dependency-based relation mining and rule-based temporal parsing together capture interactions among disease, gene, and treatment for diagnosed patients, along with clinically relevant temporal signals beyond isolated entity detection.
- c. Although annotated datasets like the NCBI Disease Corpus are often used to test NER models, this study is one of the few that shows different uses of inline annotations and syntactic relations. It also uses corpus statistics to create structured representations of clinical knowledge, expanding the potential applications of these corpora beyond model evaluation.

The rest of this article is structured as follows: Section 2 reviews related work; Section 3 introduces our methodology and experimental setup; Section 4 presents and discusses results in depth and Section 5 concludes the study.

2. Literature Review

Natural language processing (NLP) used for extracting key information from electronic health records (EHRs), to improve healthcare decision-making is being increasingly applied. Hossain et al. (2024) explored the integration of NLP with EHR systems and illustrated how models like BERT and spaCy can work process structured and unstructured data. It enhances clinical workflows and patient outcomes, especially in the identification of diagnostic and treatment patterns. But the study also pointed to ongoing challenges, such as data quality problems and the interpretability of these models. Koleck et al. (2019) conducted a review of NLP used to extract symptoms from free-text narratives

in EHRs. Their analysis applied rule-based systems and machine learning approaches. Most studies focused on disease classification rather than in-depth symptom analysis which requires further research. Ten Hoope et al. (2025) used clinical note text mining to recognize patient characteristics contained in EHRs. Their results show how NLP can convert unstructured clinical text into structured data to support patient care.

GatorTronGPT, a sizable clinical language model trained on substantial clinical and general English corpora, was created by Peng et al. in 2023. In biomedical NLP tasks, this model performs better than its general-purpose counterparts, especially when it comes to producing synthetic clinical text that supports downstream applications. Thatoi et al. (2023) investigated using natural language processing (NLP) to extract clinical data that could help with cancer prognosis. They assessed methods such as text classification, sentiment analysis, and named entity recognition and identified important obstacles like the lack of high-quality annotated datasets and linguistic variability in clinical documentation. An overview of NLP models for extracting information from medical texts was given by Chianumba et al. (2024). For tasks like entity recognition, relation extraction, and temporal analysis, they highlighted domain-specific models like BioBERT and ClinicalBERT. The review also addressed challenges including algorithmic bias, model transparency and the integration of extracted insights into clinical decision-support systems. Similarly, Raisa and Yasmin (2025) examined deep learning and machine learning techniques for Bangla Sign Language identification and highlighted the value of domain-specific datasets and sophisticated neural networks for enhancing language comprehension tasks. Furthermore, Sim et al. (2024) systematically reviewed the use of NLP to analyze unstructured patient-reported outcomes (PROs) in cancer populations, noting the clinical utility of converting unstructured data into actionable insights. They highlighted how NLP and machine learning techniques facilitate risk stratification and adverse outcome prediction.

Despite significant progress in medical NLP, current models lack sufficient context awareness for accurate relation and temporal extraction; have limited integration of semantic role labeling with temporal modeling and often focus on isolated tasks rather than unified, domain-specific pipelines for extracting structured clinical knowledge (Chianumba et al. (2024), Hossain et al. (2024), Koleck et al. (2019), Peng et al. (2023), ten Hoope et al. (2025), and Thatoi et al. (2023)).

3. Methodology

To build an end-to-end medical NLP pipeline that transforms annotated biomedical abstracts into structured knowledge by integrating NER, relation and temporal extraction, several tools and technologies have been employed.

3.1 Experimental Setup

Pipelines sequentially process biomedical abstracts by utilizing the ‘MedicalDataLoader’, which employs regular expressions to load the annotated text as a DataFrame, followed by preprocessing steps through spaCy. Exploration and analysis by ‘MedicalDataAnalyzer’ involve computing various statistics related to the document/entities and facilitating model comprehension. The named entities are recognized utilizing BioBERT (512 tokens max) and spaCy, applied for the first 100 documents. Dependency parsing in addition to keyword rules ("treatment", "drug") are used for relation extraction of 50–100 documents.

3.2 Dataset Description

This study adopts the NCBI Disease Corpus as the basic dataset for biomedical natural language processing. This dataset consists of biomedical abstracts annotated with inline XML entity tags. Each record includes a document ID, a title, and an abstract. Its explicit entity spans can support downstream tasks such as NER and relation extraction, eliminating the need for additional manual annotation.

3.3 Data Loading and Parsing

A class-specific MedicalDataLoader module converts unstructured tab-separated data entries into DataFrame format. A regular expression is used to parse and retrieve annotated entities based on the following rule:

$$<category="([^\"]+)">([^\<]+)</category>$$

This guarantees accurate span retrieval as well as retention of their semantic categories (e.g., DiseaseClass, SpecificDisease). The value that is derived from this module is the creation of a structured biomedical data set for downstream processing.

3.4 Text Preprocessing

The MedicalTextPreprocessor class uses the spaCy library and the en_core_web_sm model to preprocess text. The preprocessing pipeline follows these steps in order:

- a. Cleaning: Eliminate special characters (except for hyphens) and standardize whitespace.
- b. Tokenization: Split text into individual tokens.
- c. Stopword Removal: Eliminate frequent stopwords while retaining essential medical terminology like “cancer” and “protein”.
- d. Lemmatization: Transform tokens into their fundamental forms.

Unlike any generic NLP processing system, context-based preprocessing has been used in order to maintain biomedical semantics.

3.5 Exploratory Data Analysis

MedicalDataAnalyzer will analyze the data corpus statistically and structurally through:

- a. Document and entities distribution analysis
- b. Text length and word token distribution analysis
- c. Category visualization of entities
- d. Word cloud visualization
- e. Dependency-based matrix co-occurrence analysis

EDA is not only applied for visualizations but is also used as an analysis tool to detect entity imbalance and decide on what NER to use (BioBERT or spaCy fallback).

3.6 Named Entity Recognition

NER is carried out through BioBERTNER which is built from:

- a. dmis-lab/biobert-base-cased-v1.2 (Hugging Face Transformers)
- b. Truncated to 512 tokens per batch
- c. Entity output will comprise labels, confidence, and spans

BioBERT is utilized as it allows for more accurate biomedical entity recognition compared to regular NLP models.

3.7 Relation and Temporal Extraction

Relation and temporal extraction improve entity recognition by providing information about both how the entities relate and when the events take place. Relation extraction uses dependency parsing and keywords such as “risk” and “treatment” to extract information about entities and their connections; examples of relationships are BRCA1’s relation to breast cancer and ovarian cancer risk found in document 8589723 (“80% risk of breast cancer,” “40% risk of ovarian cancer”). The TemporalExtractor identifies time-related biomedical expressions

using regex-based rules. It extracts durations (e.g., “days”, “weeks”), frequencies (e.g., “daily”, “weekly”), and explicit age or time references (e.g., “by the age of 70”). These temporal tags are mapped with the discovered entities. For instance, in document 8589723, risk factors of breast and ovarian cancers are mapped to “by the age of 70”. It allows us to represent biomedical relations aware of time.

3.8 Pipeline Execution and Evaluation

The entire workflow demonstrated in Figure 1, is orchestrated through the main pipeline, executing the steps sequentially.

4. Results and Discussion

A total of 593 annotated biomedical abstracts were processed using the medical NLP pipeline, resulting in extraction of 4424 entities with an average of 7.46 per document.

4.1 Exploratory Data Analysis

As shown in Table 1 below, there is considerable variance in abstract length, with an average of 1,184 tokens (SD = 423). In addition, interquartile range (850-1,467 tokens) reveals the presence of notable variation in the size of documents.

Table 1: Text Length Statistics

Statistic	Value 1	Unit
Count	593	count
Mean	1183.71	characters
Standard Deviation	423.34	characters
Minimum	196	characters
25th Percentile	850	characters
50th Percentile (Median)	1178	characters
75th Percentile	1467	characters
Maximum	2591	characters

4.2 Entity Category Distribution and Properties

The bar chart in Figure 2 and the pie chart in Figure 3 demonstrate how often each annotated entity category appears in the corpus. Specific Disease has the highest frequency (about 2500 mentions) and mentions dominate at 56.7%, indicating most annotations refer to concrete, named diseases. Modifier entities (1100) are the second most common with a proportion of 24.7%, capturing qualifiers such as severity, status or attributes linked to diseases. Disease Class

appears moderately (about 700) and accounts for 16.3%, indicating more general disease group terms are also present. However, it is less frequent than specific diseases. Composite Mention is rare (around 100) and holds 2.3%, suggesting complex multi-disease mentions are annotated less often than single, specific diseases.

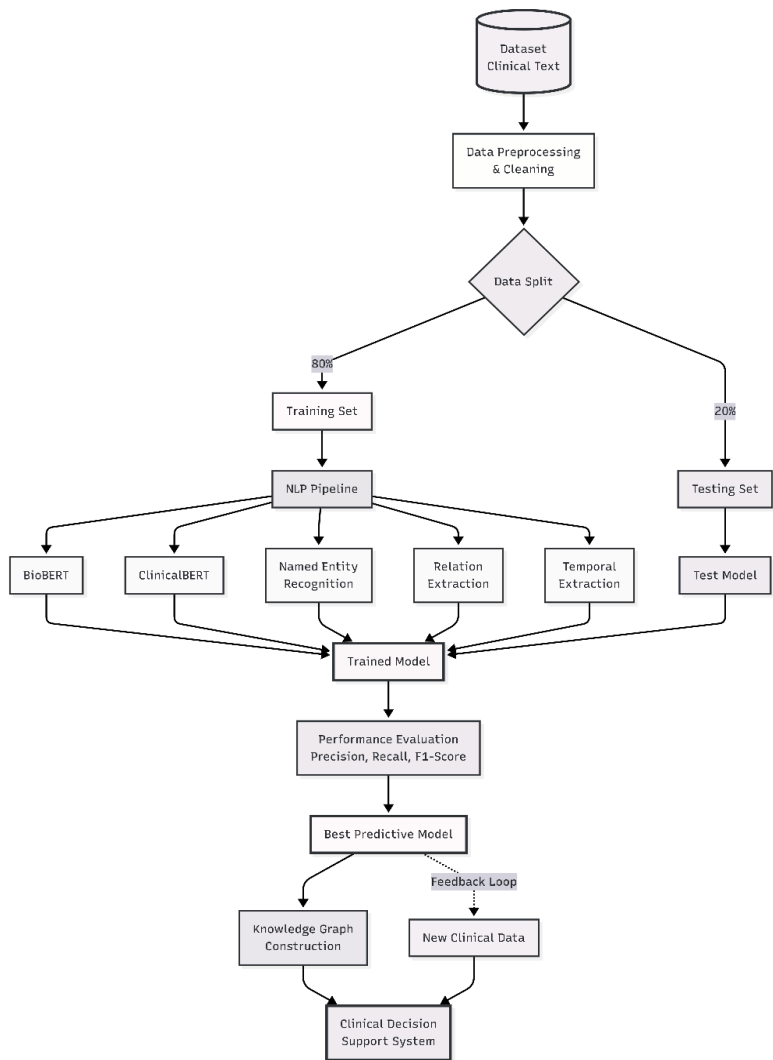


Figure 1: Proposed System workflow Diagram.

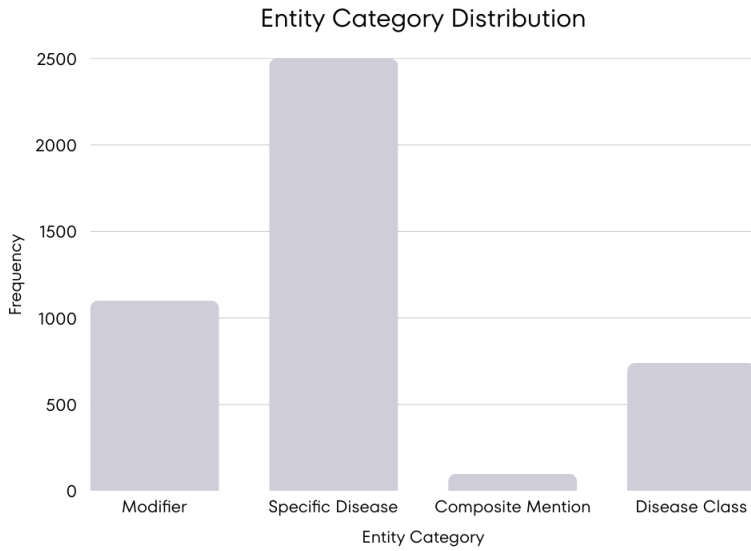


Figure 2: Entity Category Distribution.

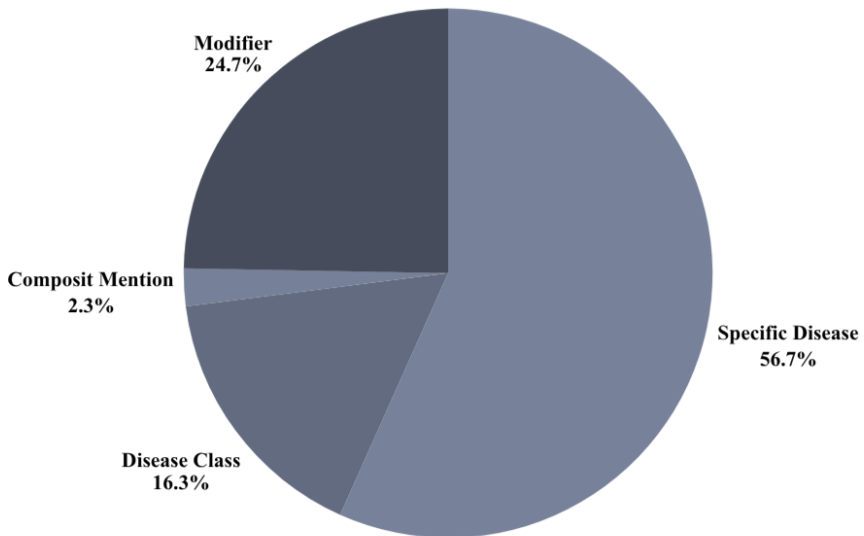


Figure 3: Entity Category Properties

4.3 Entity Count Per Document

Figure 4 illustrates the distribution of tagged entity counts per document, showing how densely each abstract is annotated with disease-related mentions. Each bar in the histogram indicates the number of documents that contain a particular range of entities. The horizontal axis shows the entity count per document, and the vertical axis shows the document frequency.

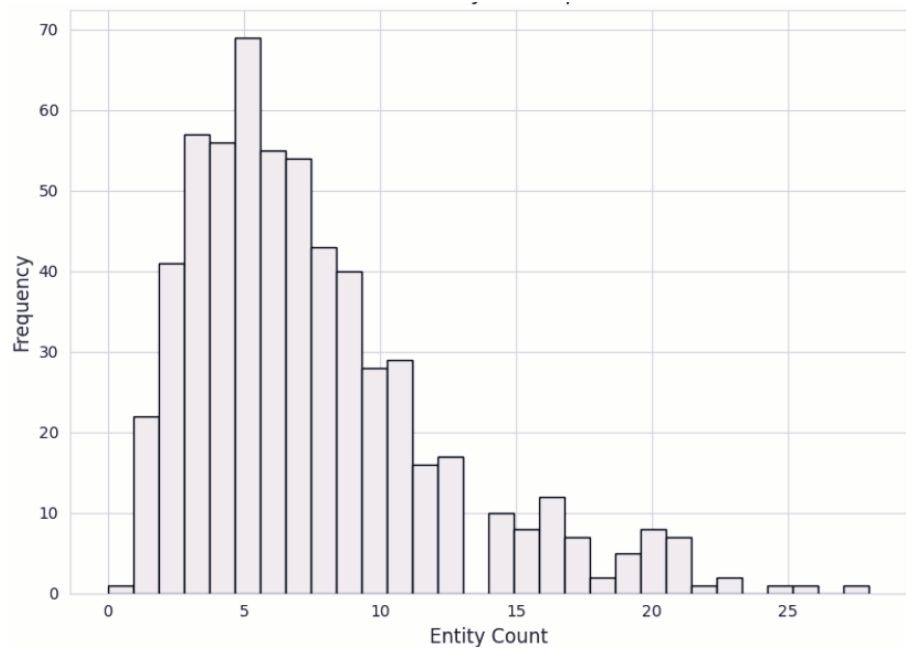


Figure 4: Entity Count Per Document

4.4 Named entity recognition with BioBERT

With an average of about 35 entity mentions per abstract, the BioBERT-based Named Entity Recognition (NER) tool identified 3,490 entities from the first 100 documents. A fairly even distribution of entity subtypes in the corpus is indicated by the model's output of two primary label categories (LABEL 1 and LABEL 0) with close frequencies (1,790 and 1,700 mentions). The model performed flawlessly on a dense cluster of biomedical entities, demonstrating its suitability for large-scale concept, even though some classifier weights were initialized randomly.

4.5 Relation and Temporal information extraction

The pipeline is advanced beyond standalone entity recognition by the relation and temporal extraction modules. A total of 613 dependency-based relations were found. Duration terms amounted to just 28 since most medical abstracts rarely use numbers for describing time. Rather, the events are described in implicit form (e.g., progression of a disease or discovery in a genetic study related to BRCA1).

4.6 Most Frequent Entities

Hereditary and oncologic conditions, such as DM, DMD, ALD, APC, and PWS medical terminologies in the dataset, are predominant, with frequencies ranging from 31 to 117 mentions, according to analysis of the top 20 most frequent entities in the Table 2. Terms like breast cancer, tumour, and entities specific to a gene or disease (like G6PD deficiency, VHL, and FRDA) appear frequently, indicating the pipeline captures important clinical concepts.

Table 2: Top 20 Most Frequent Entities

Entity	Count	Entity	Count
DM	117	Aniridia	44
DMD	104	breast cancer	39
ALD	84	FAP	38
APC	81	A-T	37
PWS	75	myotonic dystrophy	37
WAS	60	Tumor	35
G6PD deficiency	54	VHL	35
HD	51	TSD	35
PKU	51	Cancer	34
FMF	31	FRDA	31

4.7 Comparative Analysis of Biomedical NLP Pipelines

The comparison in Table 3 shows that existing pipelines either focus on entity/relation extraction or temporal modeling separately, whereas the proposed system integrates all components within a unified framework.

Table 3: Comparison of Biomedical NLP Pipelines

Feature	Perera et al. (2020)	Jolly et al. (2024)	Tu (2022)	Proposed Study
Entity Recognition	Rule + classical + partial ML	SciSpaCy + BioBERT	BiLSTM / CRF / BERT-based	BioBERT-based NER
Relation Extraction	Rule + dependency-based + ML	Not fully integrated	BERT-CNN relation model	Dependency + keyword-based
Temporal Extraction	Rule-based regex (limited)	Not included	BERT-based temporal modeling	Rule-based regex
Temporal Awareness	Limited	No	Yes (model-based)	Limited (rule-based)
Generalization	Low (dataset-specific)	Medium (multi-dataset eval)	Medium	Limited (NCBI-focused)

5. Conclusion

This work shows that a unified, domain-aware NLP pipeline can reliably convert annotated biomedical abstracts into structured clinical knowledge suitable for downstream analysis and decision support. The system integrates data parsing, spaCy-based pre-processing, BioBERT-driven NER, and dependency and rule-based relation and temporal extraction to capture entities, their interactions and their limited temporal context from 593 abstracts. Quantitative results 4,424 annotated entities, dense disease-modifier co-occurrence, an average of 35 model-detected entities per abstract, 613 extracted relations, 28 duration expressions and a set of highly frequent genetic and oncologic entities, demonstrate that the pipeline recovers clinically meaningful patterns and highlights key disease and gene hubs. These findings address gaps in prior work by moving beyond isolated tasks toward an end-to-end, interpretable framework that exposes the structure of biomedical text and can be extended with finer-grained temporal models and task-specific fine-tuning for real-world clinical decision support and predictive analytics.

5.1 Future Works

Future work will focus on strengthening temporal reasoning through the incorporation of more advanced sequence models and context-aware temporal linkers to better capture implicit and long-range time expressions. To increase the precision of disease-gene-treatment association mining in bigger biomedical corpora, the pipeline can also be expanded to employ transformer-based relation extraction models. Further, it would be desirable to apply the methodology to other datasets than the NCBI Disease Corpus, to assess its robustness and generalizability across diverse biomedical text sources. Predictive analytics and real-world clinical decision assistance may benefit from additional domain specific embedding and task-specific fine-tuning.

Acknowledgment

I would like to thank Almighty for the strength to complete this work. I thank Jasiya Fairiz Raisa for mentorship, Bangladesh University of Professionals for supporting this study, and my colleagues and family for their continuous encouragement.

Declaration of Competing Interest

The author declares that there are no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability

The data used in this study are publicly available from the NCBI Disease Corpus. No new data were generated or collected for this work.

Funding

This research received no external funding.

AI Declaration

AI tools were used only for language refinement, grammar correction, and minor coding assistance for error fixing during manuscript preparation. They were not used for research design, experimentation, or result interpretation. All analysis, implementation, and conclusions were carried out and verified by the authors.

References

- Chianumba, E. C., Ikhalea, N., Mustapha, A. Y., & Forkuo, A. Y. (2024). NLP models for extracting healthcare insights from unstructured medical text. *International Journal of Advanced Multidisciplinary Research and Studies*, 4(6), 1533–1553. <https://doi.org/10.62225/2583049X>. 2024.4.6. 4059
- Hossain, M. R., Mahabub, S., Al Masum, A., & Jahan, I. (2024). Natural language processing (NLP) in analyzing electronic health records for better decision making. *Journal of Computer Science and Technology Studies*, 6(5), 216–228. <https://doi.org/10.32996/jcsts.2024.6.5.18>
- Jolly, A., Pandey, V., Singh, I., & Sharma, N. (2024). Exploring biomedical named entity recognition via SciSpaCy and BioBERT models. *The Open Biomedical Engineering Journal*, 18, Article e18741207289680. <https://doi.org/10.2174/0118741207289680240510045617>
- Koleck, T. A., Dreisbach, C., Bourne, P. E., & Bakken, S. (2019). Natural language processing of symptoms documented in free-text narratives of electronic health records: A systematic review. *Journal of the American Medical Informatics Association*, 26(4), 364–379. <https://doi.org/10.1093/jamia/ocy173>
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234–1240. <https://doi.org/10.1093/bioinformatics/btz682>
- Mirza, J. F., Shafi, M. M. S., & Moheb, S. H. (2025). Comparison of machine learning for sentiment analysis in movie reviews. *Journal of FST*, 3(1), 147–170. <https://doi.org/10.64494/JFST/v3i1/jm/2025/07/147-170>
- National Center for Biotechnology Information. (n.d.). *Welcome to NCBI*. National Library of Medicine from <https://www.ncbi.nlm.nih.gov>. [Access Date: 5 July 2026]

- Peng, C., Yang, X., Chen, A., Smith, K. E., PourNejatian, N., Costa, A. B., Martin, C., Flores, M. G., Zhang, Y., Magoc, T., Lipori, G., Mitchell, D. A., Ospina, N. S., Ahmed, M. M., Hogan, W. R., Shenkman, E. A., Guo, Y., Bian, J., & Wu, Y. (2023). A study of generative large language model for medical research and healthcare. *npj Digital Medicine*, 6, Article 210. <https://doi.org/10.1038/s41746-023-00958-w>
- Perera, N., Dehmer, M., & Emmert-Streib, F. (2020). Named entity recognition and relation detection for biomedical information extraction. *Frontiers in Cell and Developmental Biology*, 8, Article 673. <https://doi.org/10.3389/fcell.2020.00673>
- Raisa, J. F., & Yasmin, R. (2025). Bangla sign language recognition: A comprehensive review of machine learning approaches and data sources. *Journal of FST*, 3(1), 53–86. <https://doi.org/10.64494/JFST/jr/2025/03/53-86>
- Salek Faramarzi, N., Patel, M., Bandrupally, S. H., & Banerjee, R. (2023). Context-aware medication event extraction from unstructured text. In *Proceedings of the 5th Clinical Natural Language Processing Workshop* (pp. 86–95). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.clinicalnlp-1.11>
- Sim, J. A., Huang, X., Horan, M. R., Baker, J. N., & Huang, I. C. (2024). Using natural language processing to analyze unstructured patient-reported outcomes data derived from electronic health records for cancer populations: A systematic review. *Expert Review of Pharmacoeconomics & Outcomes Research*, 24(4), 467–475. <https://doi.org/10.1080/14737167.2024.2322664>
- ten hoope, S., Welvaars, K., van Geijtenbeek, K., Klok-Everaars, M., van Schaik, S., & Karapinar-Çarkit, F. (2025). Applying text-mining to clinical notes: The identification of patient characteristics from electronic health records (EHRs). *BMC Medical Informatics and Decision Making*, 25(1), Article 302. <https://doi.org/10.1186/s12911-025-03137-x>

- Thatoi, P., Choudhary, R., Shiwlani, A., Qureshi, H. A., & Kumar, S. (2023). Natural language processing (NLP) in the extraction of clinical information from electronic health records (EHRs) for cancer prognosis. *International Journal of Membrane Science and Technology*, 10(4), 2676–2694.
- Tu, H. (2022). Extraction of temporal information from clinical free text, *Doctoral dissertation, MSc. Thesis, The University of Manchester*.