Original Article
ISSN (Online): 2582-7472

ShodhKosh: Journal of Visual and Performing Arts
June 2024 5(6), 582–594

# A RECENT SURVEY ON TEXT MINING TOOLS & TECHNIQUES IN THE BIO-MEDICAL DOMAIN

Pawan Makhija [1], Dr. Sanjay Tanwani [2]

[1] Research Scholar, Department of Computer Engineering, Institute of Engineering and Technology, DAVV, Indore, India
[2] Professor, School of Computer Science and IT, DAVV, Indore, India

**Corresponding Author**

Pawan Makhija,
pawanmakhija@acropolis.in

**DOI**
10.29121/shodhkosh.v5.i6.2024.1823

**Funding:** This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.



## ABSTRACT

Text mining in the biomedical domain has emerged as a crucial technique that capitalizes on the advancements of natural language processing and machine learning methodologies to extract valuable insights from the biomedical texts available to us from different sources. This research paper explores the methods and recent advancements in text mining techniques for the biomedical field. We analyze various approaches, including Named Entity Recognition (NER), information retrieval, and machine learning algorithms, focusing on their application to biomedical data.

Biomedical text mining aims to extract valuable information from a vast and diverse array of unstructured medical text data available from different sources, which can significantly contribute to the advancements in medical research and healthcare improvement. This research paper reflects a comprehensive exploration of fundamental concepts of biomedical textual data mining, its techniques, software, applications of biomedical textual data mining, a literature survey on clinical text mining, and its key challenges. It examines the effectiveness of these techniques in identifying and categorizing biomedical entities like genes, proteins, diseases, and drugs. Through case studies and empirical evaluations, we demonstrate how text mining contributes to knowledge discovery, improves data management, and supports decision-making in healthcare and research. This study aims to provide a detailed overview of state-of-the-art biomedical text mining, offering insights into future directions and potential improvements in this rapidly evolving field.

**Keywords:** Text Data Mining, NLP, bNER, CNN, BERT, EHRs, CRF

## 1. INTRODUCTION

Text mining in the medical domain is extracting useful information from a large corpus. Vast volumes of medical literature, coming from various sources, are common motivations for biomedical text data mining. PubMed/Medline literature is growing exponentially at the rate of publication. Biomedical literature is commonly based on unstructured data repositories, where text mining comes into play.

Text mining in biomedical informatics is a field that converges with computational linguistics, natural language processing (NLP), and bioinformatics. It mainly focuses on extracting valuable insights and knowledge from the vast amount of unstructured textual data available in the medical field through various sources, which include scientific literature, clinical notes, research articles, clinical records,

diagnostic reports, Electronic Health Record (EHRs), and multiple kinds of prescriptions.

Biomedical text mining aims to accelerate the discovery of new insights, trends, new drugs, and relationships in medicine and healthcare. The main objective of biomedical text mining is converting unstructured textual data into a structured and analyzable form of data that can be further processed for various utilities in medical research, clinical decision support, discovery of new drugs, and healthcare management. This process involves several key steps:

- Data collection and Preprocessing involve the collection of biomedical texts from different sources like PubMed, clinical databases, Electronic health records (EHRs), research articles, patients' prescriptions, and reports. Pre-processing and cleaning textual data in the medical field involves the removal of discrepancies, formatting, and extraneous information from unstructured data. Preprocessing the text data in the biomedical domain includes tasks like lowercasing, tokenization, removing special characters, stemming, lemmatization, and removing stopwords.
- Text annotation in unstructured clinical text is a process of labeling and categorizing specific elements in the text available from various sources. Biomedical named entity Recognition (bNER) targets to identify medical entities available in clinical text data.
- Information extraction focuses on identifying relationships between biomedical entities. In later stages, semantic parsing converts unstructured biomedical text data into structured text data.
- Text classification aims to classify clinical documents into relevant categories.
- Information retrieval is another stage that instructs the development of search algorithms to retrieve specific clinical documents containing keywords, concepts, or patient-related information. Then, concise summaries of lengthy clinical text documents are generated to provide quick overviews of patient cases, treatments, and medical histories.
- Biomedical information extracted from various documents is now integrated to build a structured knowledge graph. Knowledge graphs show the relationships between multiple entities and help us understand complex clinical documents comprehensively.

The steps mentioned above contribute to the overall text-mining process in the clinical domain. Using these steps, biomedical text mining enables researchers or practitioners to derive valuable insights from a vast landscape of biomedical textual data.

**Figure 1**

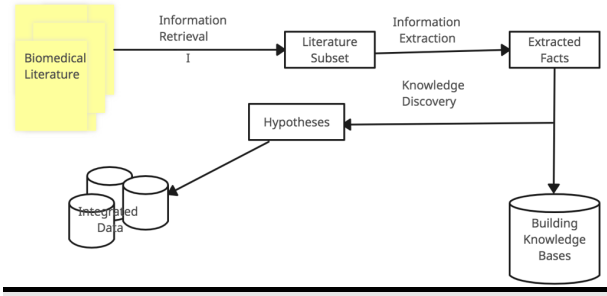


**Figure 1** Text Mining Concept

## 1.1. EVALUATION METHODOLOGY USED FOR TEXT MINING IN THE BIOMEDICAL DOMAIN

Evaluation methodologies used in text mining for biomedical data play an essential role in assessing the effectiveness and accuracy of various techniques and algorithms. To evaluate the performance of different approaches used in text mining in the medical field, this paper sheds light on the different types of metrics. The metrics are generated using a confusion matrix, where a confusion matrix is a specific matrix that aims to determine the performance of the classification model. The confusion matrix gives an idea about whether the classifier returns the correct label; alternatively, it also provides information about misclassified labels in the system.

The confusion matrix analyses the types of errors that the predictive model generates. It constitutes the detailed enumeration of several true positives (TP indicates instances that are correctly identified and returned correct labels, too), false positives (FP, the classifier assigns returns a wrong label to a data point), true negatives (TN, the model has accurately refrained from returning label that is not guaranteed), and false negatives (FN, the classifier fails to return a label that should have been returned) [18].

The performance metrics are quantitative measures that aim to analyze the effectiveness and accuracy of a text-mining model. Commonly used performance metrics are precision, which accesses the proportion of actual positive outcomes among all predicted positives; recall, which reveals the proportion of actual positives correctly identified; and F1 score, which merges precision and recall to provide a balanced evaluation of the model. Evaluation of the classification process is carried out using recall, precision, and F measures with the following formulations:

$$\text{Precision class} = TP / (TP + FP),$$

$$\text{Recall class} = TP / (TP + FN),$$

$$F = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}).$$

An inverse correlation exists between precision and recall. More precisely, when there is an enhancement in quality (precision), there is a subsequent decline in another quantity (recall), and vice versa. To achieve an optimal solution to this F1 score, a harmonic mean of the two quantities mentioned is needed. [7,17]

Accuracy is another performance metric that is used in text mining. It considers true negatives and gives the idea that the classifier returns the same output that it must return. It indicates that the text mining model correctly classifies and predicts instances in a dataset. It is a ratio of correctly predicted instances to the total number of instances:

$$\text{Accuracy} = (TP + TN)/(TP+TN+FP+FN)$$

## 2. REVIEW OF LITERATURE

Ahmad PN, Shah AM, and Lee K. propose the implementation of a Support Vector Machine (SVM) model designed for the assessment of binary relevance, and it is observed that the classifier chain configuration yielded the F1 score of 0.58, particularly when scrutinizing the tangible dimensions of the service quality evaluation framework. It has been determined that the Logistic Regression (LR)

model utilizing binary relevance methodology attained the highest recorded score among the various tested models. Concurrently, it has been noted that the Naïve Bayes (NB) model, when applied with a label power set approach, achieved the superior score for the responsiveness criterion, which was quantified at 0.63, while the Logistic Regression model utilizing a label power set methodology received the highest evaluation score for the empathy dimension, which was impressively noted at 0.88. It is essential to highlight that a consistently elevated F1 score is exclusively realized in the application of Support Vector Machines utilizing classifier chains across all dimensions of service quality, which encompass tangibility, reliability, responsiveness, assurance, and empathy, as delineated in the SERVQUAL framework.[1]

Lee J [2020] introduces a model known as BioBERT, which stands for Bidirectional Encoder Representations from Transformers designed for the Text Mining biomedical domain; this model has been pre-trained on extensive and diverse biomedical corpora that involves a wide range of biomedical literature. BioBERT performs superior to BERT and various other models in biomedical text mining tasks when it has undergone pre-training on a specialized biomedical corpus. BioBERT model has shown improvements in three biomedical text mining tasks, namely biomedical named entity recognition, which shows an enhancement of 0.62% in F1 score, biomedical relation extraction, achieving an improvement of 2.80% in F1 score, and biomedical question answering exhibits an enhancement of 12.24% in mean reciprocal rank (MRR). Their comprehensive study indicated that pre-training BERT on biomedical corpus considerably augments its capability to comprehend and navigate the complexities inherent in biomedical texts, leading to a more profound understanding and interpretation of such specialized information [2].

Houssein EH, Mohamed RE, and Ali AA proposed a methodology involving the utilization of stacked embeddings to enhance the findings of prior investigations about the i2b2 2014 challenge. The dataset associated with the i2b2 heart disease risk factors challenge has demonstrated considerable advancement by applying stacked embeddings, which serves as a mechanism for amalgamating multiple embeddings. By employing a combination of BERT and character embeddings (termed CHARACTER-BERT Embedding), an F1-score of 93.66% is achieved on the test dataset.[4]

Shahab Shamshirband, Mahdis Fathi, Abdollah Dehzangi, Anthony Theodore Chronopoulos, and Hamid Alinejad-Rokny (2021) introduced a hybrid approach that integrated ensemble techniques based on deep learning. This hybrid methodology yielded better accuracy outcomes than singular methods. Techniques that are based on deep learning, like Convolutional Neural Networks (CNN), Deep Belief Networks (DBN), and autoencoders, have been evidenced as formidable instruments in the fields of disease detection, encompassing preprocessing, feature extraction, feature selection, classification, and clustering phases.[10]

Xiangyang Li, Huan Zhang, and Xiao-Hua Zhou acquired a pre-trained BERT model on the enormous Chinese clinical corpora. The proposed model starts with the comparative analysis between the baseline model and the meticulously fine-tuned version of the BERT model. Following this comparison, a thorough evaluation process involving the involvement of Long Short-Term Memory (LSTM) networks and Conditional Random Fields (CRF) layers resulted in an F1 score of 89.56%.[11]

Iñigo Jauregi Unanue, Ehsan Zare Borzeshi, and Massimo Piccardi put forth the investigation into the comparative analysis of two distinct architectures of Recurrent Neural Networks, which are categorized explicitly as Bidirectional Long

Short-Term Memory (LSTM) networks and Bidirectional LSTM combined with Conditional Random Fields (CRF), with a particular focus on their utility for the task of information extraction. They conducted a rigorous comparison of these advanced neural network models against the traditional baseline model based on Conditional Random Fields (CRF).[12]

Jiho Noh B and Ramakanth Kavuluru, in the year 2021, proposed a comprehensive and innovative model that aims to enhance the quality and efficacy of biomedical word embeddings by utilizing the advanced capabilities of the BERT model, which stands for Bidirectional Encoder Representations from Transformers. This scholarly article sheds considerable light on the intricate process of identifying words that possess identical meanings or are closely related synonyms within the specialized context of the biomedical domain, thereby contributing to the ongoing discourse in this field.[13]

Brincat, A., and Hofmann 2022 put forth a comprehensive model that meticulously elucidated the intricate text mining pipeline, which is specifically designed to analyze and synthesize information from the vast corpus of biomedical literature, ultimately yielding a list of genes that have the potential to confer resistance to antibiotics, thereby contributing to an enhanced understanding of this critical issue in the field of microbiology and pharmacology.[19]

This literature review showcases the diverse and innovative approaches being explored in biomedical text mining. From advanced machine learning techniques to specialized language models, these studies collectively enhance the accuracy, efficiency, and applicability of text mining in the biomedical domain. The ongoing research in this field promises to refine further our ability to extract valuable insights from the vast corpus of biomedical literature, potentially revolutionizing healthcare research and clinical practice.

**Table 1**

**Table 1 Summary of Literature on Biomedical Text Mining Techniques and Models**

| S.no | Title | Technique | Year |
|---|---|---|---|
| 1 | A Review on Electronic Health Record Text-Mining for Biomedical Name Entity Recognition in Healthcare Domain by Smith et al. (2023). | SVM | 2023 |
| 2 | BioBERT: a pre-trained biomedical language representation model for biomedical text mining by Lee et al. (2020). | BIOBERT | 2020 |
| 3 | Literature mining for context-specific molecular relations using multimodal representations (COMMODAR) by Zhang & Zhao et al. (2020) | CNN | 2020 |
| 4 | Heart disease risk factors detection from electronic health records using advanced NLP and deep learning techniques by Wang & Li et al. (2023) | CHARACTER-BERT Embedding | 2023 |
| 5 | A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues by Patel & Kumar et al. (2021). | Hybrid & ensemble methods based on deep learning | 2021 |
| 6 | Chinese clinical named entity recognition with variant neural structures based on BERT methods by Liu et al. (2020) | BERT model | 2020 |
| 7 | Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition by Jones et al. (2017). | RNN | 2017 |
| 8 | Improved biomedical word embeddings in the transformer era by Chen et al. (2021). | BERT model | 2021 |

| | | | |
|---|---|---|---|
| 9 | Automated extraction of genes associated with antibiotic resistance from the biomedical literature by Johnson et al. (2022). | NLP | 2022 |
| 10 | Implementation of a Support Vector Machine (SVM) Model for Binary Relevance Assessment \|by Ahmad et al. (2023) \| | SVM | 2023 |
| 11 | BioBERT: Bidirectional Encoder Representations from Transformers for Biomedical Text Mining by Lee et al. (2020) \| | BioBERT | 2020 |
| 12 | Stacked Embeddings for i2b2 Heart Disease Risk Factors Challenge by Houssein et al. (2023) \| | CHARACTER-BERT Embedding | 2023 |
| 13 | Hybrid Approach Integrating Ensemble Techniques Based on Deep Learning for Disease Detection \|by Shamshirband et al. (2021) \| | Hybrid & ensemble methods | 2021 |
| 14 | Pre-trained BERT Model on Chinese Clinical Corpora for Named Entity Recognition by Li et al. (2020) \| | BERT model | 2020 |
| 15 | Comparative Analysis of Recurrent Neural Networks for Information Extraction by Unanue et al. (2017) \| | RNN | 2017 |
| 16 | Enhancing Biomedical Word Embeddings Using BERT for Synonym Identification by Noh & Kavuluru et al. (2021) \| | BERT model | 2021 |
| 17 | Text Mining Pipeline for Analyzing Biomedical Literature on Antibiotic Resistance by Brincat & Hofmann et al. (2022) | NLP | 2022 |

## 3. ADVANCED TEXT MINING APPROACHES IN BIOMEDICAL INFORMATICS

### 1) Information Retrieval

Information retrieval in the biomedical domain refers to obtaining relevant and accurate information from large volumes of healthcare-related data. The preliminary phase of text mining in the biomedical discipline is to retrieve textual resources from a large corpus available in different forms.[15] It helps researchers, clinicians, and practitioners access information from various biomedical resources like relevant articles, research papers, medical studies, or scientific databases like PubMed, Embase, Medline, and Web of Science. Some essential Information Retrieval techniques in the biomedical domain are:

### 2) NLP-driven Text Summarization

Text Summarization refers to the technique of producing concise and coherent summaries that capture the essential information from a large biomedical text document.

Biomedical literature often contains lengthy research articles from various sources that create a need to extract fruitful information.

There are two approaches for biomedical text summarization: extractive and abstractive. Keyword extraction and abstractive summarization transform biomedical literature into a concise summary.[14] In extractive summarization, we identify essential sentences or phrases directly from the original clinical text and extract their main ideas. The extraction process involves ranking the sentences based on features like word frequency, importance scores, or statistical measures and then choosing the top-ranked sentences to generate a summary.

**Figure 2**

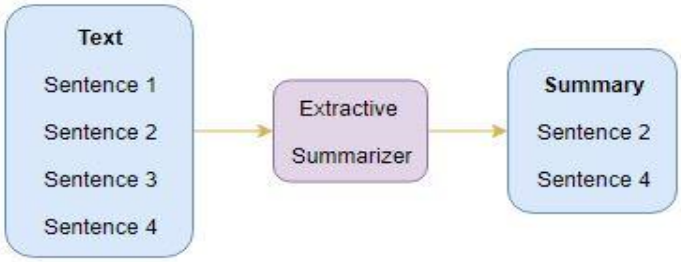


**Figure 2** Extractive Summarization

Meanwhile, abstractive summarization allows some flexibility and creativity in clinical text summarization. It generates new sentences that do not exist in the original text to create a summary of a desired text.

**Figure 3**

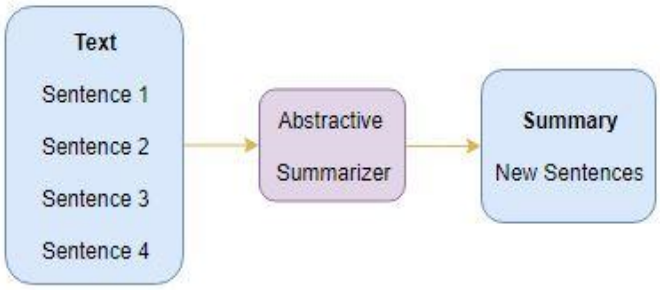


**Figure 3** Abstractive Summarization

### 3) Knowledge Extraction

Knowledge extraction is mining meaningful words from various sources from a clinical document. It is the automated retrieval of the most helpful information related to a selected topic from bodies of unstructured clinical text. In information extraction, the text files containing the unstructured data are changed to structured data through mining techniques to get valuable insights that help decision-making and contribute to new research in the clinical domain. To achieve this goal of extracting information, we develop new features from the existing ones by parsing or combining two or more features based on mathematical computation.

Named Entity Recognition (NER): Biomedical Named Entity Recognition(bNER) focuses on identifying and categorizing specific biomedical entities in unstructured data, like scientific literature, patents, clinical records, and biomedical research articles. Biomedical informatics involves analyzing and identifying bNER and healthcare-related entities in electronic health records(EHRs), such as medications, proteins, gene compounds, and diseases in unstructured medical texts [1].

The Biomedical Named Entity Recognition technique is mainly categorized under three major classes: dictionary-based approaches, rule-based approaches, and machine learning approaches [8]. From already available research, the dictionary-based approach is prone to miss undefined terms not mentioned in the biomedical dictionary. Secondly, rule-based approaches entail the specific protocols that could determine biomedical elements from biomedical textual data, and it is

seen that the subsequent regulations are often only effective in some cases. Thirdly, machine learning approaches require standard annotated training data sets. Most of the machine learning approaches are data-driven. Metrics commonly used with machine learning approaches are precision, recall, and F1 score, often used to evaluate the recognition result.[5]

**Figure 4**

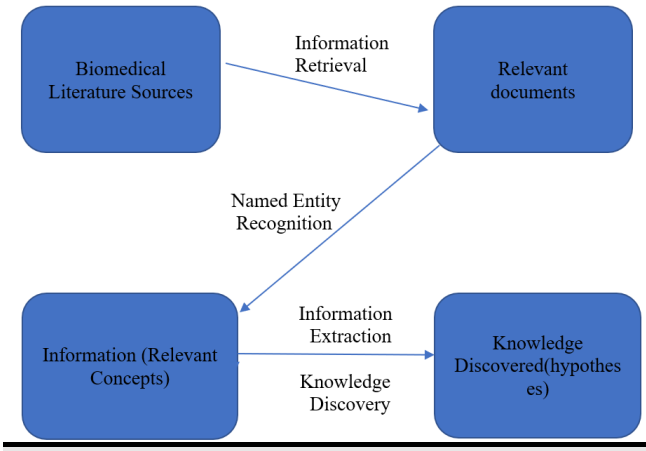


**Figure 4** Text Mining Process in the Biomedical Domain

Various data sources are available in the public domain that contain the source of knowledge that can be accessed for mining the rich text from the biomedical domain.

**Table 2**

**Table 2 Biomedical Databases and Tools with URLs**

| S.no | Name | URL |
|---|---|---|
| 1 | MEDLINE | https://www.nlm.nih.gov/databases/download/pubmed_medline.html |
| 2 | MeSH | http://www.nlm.nih.gov/mesh/meshhome.html |
| 3 | UMLS | http://www.nlm.nih.gov/research/umls |
| 4 | SNOMED | http://www.nlm.nih.gov/snomed / |
| 5 | SemMedDB | https://skr3.nlm.nih.gov/SemMedDB/dbinfo.html |
| 6 | HUGO | http://www.hugo-international.org |
| 7 | PhenoGo | http://www.lussiergroup.org |
| 8 | NLP Tools SemRep | https://semrep.nlm.nih.gov/ |
| 9 | MetaMap | https://metamap.nlm.nih.gov/ |

### 4) Text Clustering

Text clustering in the biomedical domain groups most similar biomedical text, such as medications, research articles, biomedical notes, and prescriptions, into clusters based on their content and characteristics. Text Clustering is an unsupervised learning technique that can be used to analyze unstructured and unlabeled biomedical data. The benefit of text clustering is that the clinical text files will be in multiple sub-topics, which makes it safer for essential documents to get erased from search. The clustering technique separates records in a dataset into groups so that themes in a cluster are the same while themes between the clusters differ. Acquiring the group with some value about the difficulty being addressed is

the main aim of cluster analysis. Text clustering in the biomedical domain involves K-means clustering, hierarchical clustering, density-based clustering, and topic modeling approaches like Latent Dirichlet Allocation(LDA).

Critical steps in Text clustering:

1) **Data collection & text preprocessing:** Text preprocessing in biomedical research involves tokenization, stop word removal, stemming or lemmatization, and converting biomedical text to a numerical representation suitable for clustering algorithms.
2) **Feature Extraction:** A Few of the standard techniques in text clustering in the biomedical domain include Bag-of-words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF) to assign weights to terms and create feature vectors
3) **Clustering algorithms:** K-means clustering is the simplest one. The portioning clustering method partitions the biomedical text into random segments. Hierarchical clustering builds the tree-like structures of clusters, which can take two forms: agglomerative and divisive. Density-based clustering is another clustering algorithm that groups biomedical text based on its density in the feature space. Typically, the k-means clustering algorithm operates more quickly than the hierarchical clustering algorithm [6].
4) **Similarity Measurement:** In text clustering, similarity or distance measure is applied to calculate the similarities between biomedical texts. In this regard, cosine similarity and Euclidean distance are commonly used metrics for text clustering in the biomedical domain.
5) **Text Categorization**

In text classification, the fundamental themes of biomedical literature are identified. This is achieved by allocating the texts into a collection of categories or classes based on their content, which is predefined within it. The classified document can be regarded as a 'bag of words.' Information extraction endeavors to process the actual information, whereas classification does not strive to process the actual data. In this, the terms of the biomedical literature are enumerated by the classification process. Subsequently, utilizing these counts, they discern the pivotal subjects of the text.

Text classification in the biomedical domain is contingent upon the lexicon for which the subjects are predetermined, and associations are discerned by exploring extensive, specific, synonymous, and pertinent terminology. Text categorization in the biomedical domain may involve supervised machine-learning approaches. This paper reviews documents that reflect different algorithms for text mining in a biomedical domain, like support vector machine (SVM), Naïve Bias, Convolutional neural networks (CNNs), and transformers. The main objective of text categorization is to classify and organize large volumes of unstructured biomedical text data into a fixed quantity of predefined groups, making it easier to navigate, retrieve information, and perform analysis. Text categorization in biomedical informatics has led to the classification of medical literature that can classify research articles into different fields like oncology, cardiology, neurology, etc. In the biomedical domain, text categorization can help in other areas like drug categorization, medical imaging classification, symptom classification, disease classification, bioinformatics data classification, etc. [20]

## 4. BIOMEDICAL TEXT MINING SOFTWARE

Biomedical text mining plays a crucial role in extracting valuable information from vast amounts of research data available in sources such as PubMed and Medline. This section provides an overview of various software tools and web applications designed to facilitate text mining in the biomedical domain.

**Table 3**

**Table 3 Web Applications for Biomedical Text Mining**

| S. no | Web application | Description |
|---|---|---|
| 1 | PubTator | PubTator is a text-mining tool for annotating PubMed articles with key biological entities (e.g., genes and diseases). It is available through both Web and API access. |
| 2 | LitVar | LitVar is a semantic search engine linking genomic variant data in PubMed and PMC. |
| 3 | LitCovid | LitCovid is a literature hub for tracking up-to-date scientific information about the 2019 novel Coronavirus, first created in February 2020. |
| 4 | LitSuggest | LitSuggest is a web-based literature triage and document classification system using AI and machine learning. |
| 5 | LitSense | LitSense helps make sense of the biomedical literature at the sentence level by finding the best-matching sentences given a query via a cutting-edge neural embedding approach. |
| 6 | TeamTat | TeamTat is a web-based text annotation tool for biomedical text and beyond. |

**Table 4**

**Table 4 Software Tools for Biomedical Named Entity Recognition and Text Processing**

| S. no | Software | Description |
|---|---|---|
| 1 | TaggerOne (All-purpose tagger) | TaggerOne is a general toolkit for biomedical named entity recognition and normalization. As a machine learning system, it is not entity-specific but does require training data. |
| 2 | tmChem (chemical tagger) | tmChem is an open-source software tool for identifying chemical names. tmChem achieved the highest Performance in BioCreative CHEMDNER task (over 87% F-measure) |
| 3 | DNorm (disease tagger) | DNorm is the first technique to use machine learning to recognize and normalize disease names in biomedical text. |
| 4 | GNormPlus (gene tagger) | GNormPlus is an end-to-end system that handles gene/protein name and identifier detection in biomedical literature, including gene/protein mentions, family names, and domain names. |
| 5 | SR4GN (species tagger) | SR4GN is an open-source species recognition/disambiguation tool optimized for the Gene Normalization task. |
| 6 | tmVar (mutation tagger) | tmVar extracts sequence variants in protein and gene levels (e.g., substitution, deletion, etc.) in HGVS formats. |
| 7 | SimConcept (text simplification) | SimConcept uses patterns to identify individual mentions from a composite named entity (e.g., SMAD 1, 5, and 8). |
| 8 | NegBio | NegBio is a high-performance tool for negation and uncertainty detection in clinical text (e.g., radiology reports). |

## 5. APPLICATIONS OF TEXT MINING IN BIOMEDICAL DATA

### 1) Literature review and knowledge discovery

Many techniques and tools of text mining in the biomedical domain are used to set up patterns and trends from journals and their proceedings from massive amounts of sources. To do research in the biomedical domain, we can get great sources of information from patents, scientific articles, and medical reports for chemical entities and diseases [7]. In the primary stages of knowledge discovery, recognition is done to identify words and phrases that belong to certain classes, like proteins and genes. It then determines the domain of the biomedical text. The metadata from this phase can be put into the graph database for further analysis. Biomedical text available through PubMed [7] is converted into XML format, and the machine learning module performs the bNER to diseases and chemical classes in the article. Then, this metadata is used to construct a graph database [7].

### 2) Drug Discovery

Drug discovery is a process that can potentially identify and develop new therapeutic compounds to treat different diseases. New drugs or medicines can be found with the help of text-mining techniques in the biomedical domain. Drug repositioning identifies and develops different uses for a particular drug initially designed for a specific disease.[15]

### 3) Clinical Decision Support

Clinical decision support is a technology-based process that can help healthcare professionals make decisions about patient care. It consists of descriptive or predictive analysis [16]. It considers using software that provides clinical knowledge to help inform decisions about a patient's care. An analysis of clinical decision support is aimed mainly at patients with diabetes, cancer, and cardiovascular disease and patients in the internal care unit (ICU) [1].

### 4) Biomarker Discovery

Biomarkers play a significant role in disease diagnosis, prognosis, and treatment monitoring for a particular disease. Biomarker discovery is a process in medical research and healthcare systems. Biomarkers are the measurable units by which we can access an individual's physiological, pathological, and pharmacological methods. Biomarkers can be classified into diagnostic, prognostic, or predictive categories. These can be used to select patients from a population sample for clinical trials or monitor patients' response and treatment efficacy. Biomarkers help us determine what drugs are most likely to be safely tolerated in human beings based on specific trials.

### 5) Gene Function Prediction

Gene function prediction aims at a multidisciplinary field that combines various computational methods, biological knowledge, and experimental validation. Understanding the multiple functions of genes is essential for unraveling complex molecular mechanisms in biological processes and diseases. Deep learning models were used to determine a solution for predicting gene function that can cause resistance to particular types of antibiotics.[19]

### 6) Disease Surveillance

Text mining in the biomedical domain can be used to detect disease outbreaks early by analyzing textual data available from different sources, such as news articles, social media, reports, prescriptions, electronic medical records, and others.

Disease surveillance involves interpreting and disseminating data on diseases in a defined population to reduce morbidity and mortality for public health.

## 6. CONCLUSION

Biomedical text mining has become an essential technique, leveraging natural language processing and machine learning to derive meaningful insights from extensive biomedical texts. This paper provides a comprehensive overview of current methods and developments in this field, focusing on techniques such as Named Entity Recognition (NER), information retrieval, and machine learning algorithms tailored for biomedical data.

Our analysis reveals that text mining is crucial for extracting valuable information from diverse and unstructured medical texts, contributing significantly to medical research and healthcare. The paper covers fundamental concepts, explores various techniques and software tools, and reviews the application of text mining in identifying key biomedical entities like genes, proteins, diseases, and drugs.

By presenting case studies and empirical evaluations, we highlight the impact of text mining on knowledge discovery, data management, and decision-making in healthcare. This study not only underscores the importance of these techniques but also points to future directions for further advancement. As the field evolves, ongoing improvements and innovations in text mining methods will continue to enhance our ability to manage and interpret biomedical data, ultimately supporting progress in research and improving patient care.

## CONFLICT OF INTERESTS

None.

## ACKNOWLEDGMENTS

None.

## REFERENCES

Ahmad, P. N., Shah, A. M., & Lee, K. (2023). A Review on Electronic Health Record Text-Mining for Biomedical Named Entity Recognition in the Healthcare domain. *Healthcare (Basel)*, *11*(9), 1268.

Ee, 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.

Ee, J., Lee, D., & Lee, K. H. (2020). Literature Mining for Context-Specific Molecular Relations Using Multimodal Representations (COMMODAR). *BMC Bioinformatics*, *21*(S5).

oussein, E. H., Mohamed, R. E., & Ali, A. A. (2023). Heart Disease Risk Factors Detection from Electronic Health Records Using Advanced NLP and Deep Learning Techniques. *Scientific Reports*, *13*(1), 7173.

Hu, F., Patumcharoenpol, P., Zhang, C., Yang, Y., Chan, J., Meechai, A., Vongsangnak, W., & Shen, B. (2013). Biomedical Text Mining and its Applications in Cancer Research. *Journal of Biomedical Informatics*, *46*(2), 200-210.

enganathan, V. (2017). Text Mining in Biomedical Domain with Emphasis on Document Clustering. *Healthcare Informatics Research*, *23*(3), 141-150.

oylin, A. B., & Victor, N. (2017). Knowledge Discovery in bIomedical Literature. *Research Journal of Pharmacy and Technology*, *10*(6), 1911-1918.

Ai, S., Ding, Y., Zhang, Z., Zuo, W., Huang, X., & Zhu, S. (2019). GrantExtractor: Accurate Grant Support Information Extraction from Biomedical Full-Text Based on Bi-LSTM-CRF. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1-1.

opalakrishnan, V., Jha, K., Jin, W., & Zhang, A. (2019). A Survey on Literature-Based Discovery Approaches in the Biomedical Domain. *Journal of Biomedical Informatics*, *93*, 103141.

Shamshirband, S., Fathi, M., Dehzangi, A., Chronopoulos, A. T., & Alinejad-Rokny, H. (2021). A Review on Deep Learning Approaches in Healthcare systems : Taxonomies, challenges, and open issues. *Journal of Biomedical Informatics*, *113*, 103627.

Li, X., Zhang, H., & Zhou, X.-H. (2020). Chinese Clinical nAmed Entity Recognition with Variant Neural Structures Based on BERT Methods. *Journal of Biomedical Informatics*, *107*, 103422.

Jauregi Unanue, I., Zare Borzeshi, E., & Piccardi, M. (2017). Recurrent Neural Networks with Specialized Word eMbeddings for hEalth-Domain Named-Entity Recognition. *Journal of Biomedical Informatics*, *76*, 102-109.

Noh, J., & Kavuluru, R. (2021). Improved Biomedical Word Embeddings in the Transformer Era. *Journal of Biomedical Informatics*, *120*, 103867.

Noh, J., & Kavuluru, R. (2020). Literature Retrieval for Precision Medicine with Neural Matching and Faceted Summarization. In *Findings of the Association for Computational Linguistics : EMNLP 2020* (pp. 3389-3399).

Fleuren, W. W. M., & Alkema, W. (2015). Application of Text Mining in the Biomedical Domain. *Methods*, *74*, 97-106.

Islam, M. S., Hasan, M. M., Wang, X., Germack, H. D., & Noor-E-Alam, M. (2018). A Systematic Review on Healthcare Analytics : Application and theoretical Perspective of Data Mining. *Healthcare (Basel)*, *6*(2), 54.

Solarte-Pabón, O., Montenegro, O., García-Barragán, A., Torrente, M., Provencio, M., Menasalvas, E., & Robles, V. (2023). Transformers for Extracting Breast Cancer Information from SPanish Clinical Narratives. *Artificial Intelligence in Medicine*, *143*, 102625.

Li, X., Yuan, W., Peng, D., Mei, Q., & Wang, Y. (2022). When BERT Meets Bilbo : A Learning Curve Analysis of pRe-Trained Language Model on Disease Classification. *BMC Medical Informatics and Decision Making*, *21*(Suppl 9), 377.

Brincat, A., & Hofmann, M. (2022). Automated Extraction of Genes Associated with Antibiotic Resistance from the Biomedical Literature. *Database (Oxford)*, *2022*(2022), baab077.

Wu, S., Roberts, K., Datta, S., Du, J., Ji, Z., Si, Y., Soni, S., Wang, Q., Wei, Q., Xiang, Y., Zhao, B., & Xu, H. (2020). Deep Learning in Clinical Natural language processing : A Systematic Review. *Journal of the American Medical Informatics Association*, *27*(3), 457-470.