Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 27 (2024), No. 2, pp. 383–393
DOI : 10.47974/JSMS-1263

Natural language processing for drug information extraction : Advancing knowledge discovery in biomedical literature

A. A. Koparde *
Department of Pharmaceutical Chemistry
Krishna Institute of Pharmacy
Krishna Vishwa Vidyapeeth (Deemed to be University)
Karad
Maharashtra
India

Pradnya A. Jadhav †
Department of Electronics Engineering
Yeshwantrao Chavan College of Engineering
Nagpur
Maharashtra
India

G. Jisha Annie §
Department of Pharmacy Practice
Krishna Institute of Pharmacy
Krishna Vishwa Vidyapeeth (Deemed to be University)
Karad
Maharashtra
India

Dinesh Goyal ‡
Department of Computer Engineering
Poornima Institute of Engineering & Technology
Jaipur
Rajasthan
India

---

* *E-mail:* akshadakakade@yahoo.com
† *E-mail:* ja_pradnya@yahoo.com
§ *E-mail:* jishaannie22@gmail.com
‡ *E-mail:* dinesh.goyal@poornima.org

© Taru Publications

Bhalchandra M. Hardas [@]
Department of Electronics and Computer Science
Shri Ramdeobaba College of Engineering and Management
Nagpur
Maharashtra
India

Purva Mange [#]
Symbiosis School of Planning Architecture and Design
Symbiosis International University
Maharashtra
India

Abstract

Natural language processing (NLP) has emerged as an important tool in the biomedical industry for bettering medication information extraction and other knowledge gathering. In this study, we explore the use of natural language processing (NLP) techniques to glean useful insights from massive volumes of biological literature, with the goal of better comprehending data pertaining to drugs. We want to automate the extraction of crucial aspects such drug interactions, side effects, and efficacy by utilizing sophisticated language models and semantic analysis. Effective and comprehensive drug information retrieval will be encouraged. Our innovation improves the drug development process by facilitating the rapid exploration of big databases for previously unknown connections and patterns. By facilitating the synthesis of data from several sources, NLP in biomedical research speeds up information extraction, which in turn accelerates drug development and improves patient care by allowing for evidence-based decision making. This research demonstrates the promising potential of natural language processing (NLP) for unraveling the complexities of drug-related data, ushering in a new era of learning in the biomedical field.

Subject Classification: *68N15.*

Keywords: *Knowledge discovery, Drug information, NLP, Machine learning, Feature extraction.*

1. Introduction

NLP solves [1] this problem by automating the extraction procedure, which makes it possible to analyze a variety of sources in an organized manner. NLP improves the accuracy and scope of information retrieval by removing noise and focusing on important ideas. In order to acknowledge the complex web of relationships that shapes our understanding of

[@] *E-mail:* **hardasbm@rknec.edu**
[#] *E-mail:* **purva.mange@gmail.com**

therapeutic interventions, this study focuses on the intricate interactions between diseases and therapies within the biological setting [3]. The article primarily focuses on four important steps in the preprocessing of medical texts: splitting, tokenization, part-of-speech tagging, and semantic annotation. It [2] critically examines the role of NLP in this process. Together, these phases help to refine raw data, making it possible to describe biological concepts in a more structured and detailed manner. We illuminated the potential of natural language processing (NLP) to transform knowledge discovery in the biomedical field by clarifying these fundamental stages and laying the groundwork for a thorough investigation of NLP's influence on medication information extraction [4]. Our goal is to push the boundaries of drug discovery and healthcare by contributing to the current discussion about the use of cutting-edge technology in biomedical research.

2. Drug-Disease Relation Extraction

A crucial [5] component of biomedical information retrieval is drug-disease relation extraction, which is necessary to comprehend the complex relationships between pharmacological treatments and particular medical disorders. Using sophisticated Natural Language Processing (NLP) tools, this methodology examines large datasets to [6] identify relevant drug-disease connections from biomedical publications. NLP algorithms are essential in this situation for interpreting the complex language found in medical publications. There are usually several steps in the extraction process, such as tokenization, semantic analysis, and text segmentation [7]. Through these steps, the algorithm can recognize and classify the linkages between medications and illnesses, picking up on both explicit and implicit correlations found in the text.

The potential [8] for several applications, such as drug repurposing, adverse event prediction, and personalized medicine, is enormous when successful medication-disease connection extraction is achieved. Researchers and medical practitioners can learn more about possible adverse effects, new treatment approaches, and the efficacy of particular medications for specific diseases by methodically discovering and cataloguing these correlations. The [9] intricacy and diversity of biomedical language present difficulties, but continuous improvements in natural language processing (NLP) models like contextual embedding and deep learning techniques keep improving the precision and effectiveness of drug-disease link extraction. By using technology to unlock important

knowledge hidden in biomedical literature, our research helps to further the larger objective of using technology to support better decision-making in drug development, clinical practice, and public health initiatives.

3. Proposed Method

This study's methodology resulted in a thorough approach because it was informed by earlier research [10] on text mining, machine learning, and relation extraction from medical literature. Preprocessing, the initial step, began with sentences in free-text and worked to refine and organize the raw material. The system reduced noise and created a set of annotated words through this preprocessing, which also set the stage for further research. This first phase had a crucial role in improving the data quality and enabling a more sophisticated comprehension of the drug-disease correlations.

The second part dealt with feature extraction that is, finding different features in phrases that would be important markers for relation extraction that followed. These characteristics, which were retrieved using sophisticated methods, provided contextual information that was vital for determining the complex relationships between medications and illnesses in the biomedical setting. The application of machine learning was the third and final component. A robust learning algorithm was given the data from the feature extraction phase. This machine learning [11]

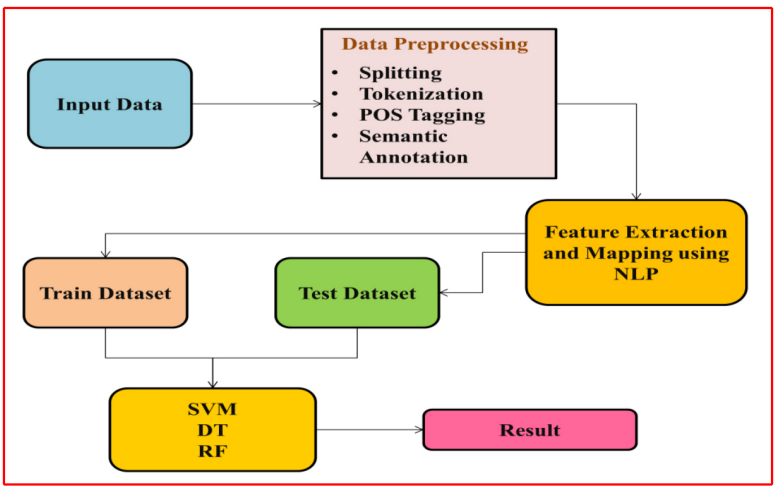

Figure 1

Systematic overview of proposed model

component finished identifying correlations between drug and disease entities by utilizing contextual characteristics and annotated data. The seamless integration of these three components is visually represented by the holistic architecture of the NLP method, as seen in Figure 1. With its multifaceted approach, the suggested methodology not only addressed the difficulties associated with relation extraction from biomedical literature but also demonstrated a methodical and efficient manner to unearth important insights within this intricate field.

This work advances the field of drug-disease relation extraction with NLP model, which has potential uses in clinical decision assistance, drug development, and biomedical knowledge discovery.

3.1 *Preprocessing*

Our methodology starts with essential Natural Language Processing (NLP) preprocessing, wherein unprocessed medical texts are refined to extract pertinent biomedical information and reduce noise. To improve the quality and relevance of the collected data, a four-stage pipeline is used: splitting for segmentation, tokenization for in-depth semantic analysis, part-of-speech tagging for grammatical context identification, and semantic annotation for nuanced representation.

3.2 *Feature Extraction*

To fully capture the subtleties of drug-disease connections in biological texts, we utilized a multimodal approach in the feature extraction step of the suggested methodology. Four different categories of features were extracted throughout this process: frequency features, lexical features, syntactic features, and semantic features.

3.2.1 Frequency Features:

A quantitative perspective on the connections between medications and illnesses in biomedical language was offered by the frequency aspects in our research. The characteristics that were included in this were the quantity of named entities (NEs) that denoted drugs and diseases, the number of nouns that denoted drugs and diseases independently, and the frequency of verbs that established connections between these entities. The features also included a word count, lemma count, and a Bag-of-Words representation that recorded the word occurrences and their associations inside the sentence [12]. This comprehensive approach provided a nuanced comprehension of the complex relationships between

medications and disorders in biological texts by taking into account not only the structural and linguistic features but also the quantification of crucial element presence. In particular, the Bag-of-Words function summarized word clusters inside phrases, describing word occurrences and facilitating a thorough examination of the text.

3.2.2 Lexical Features

An organized method for examining the word distribution in a phrase is provided by the application of bag-of-words models within lexical features. The frequency of the words is retained in this model, but their order is ignored. The bag-of-words format adds a quantitative layer to the lexical analysis by measuring the occurrence of specific phrases connected to pharmaceuticals and disorders, which helps identify important terms that contribute to the overall semantic context. They make it possible for the model to identify language linkages, patterns, and contextual meanings that are essential for figuring out how drugs and diseases are related. By combining these lexical properties with additional feature categories, the technique produces a detailed and complex representation of the text, which serves as a foundation for the succeeding steps in the extraction of the drug-disease relationship. The addition of lexical features enhances the model's comprehension of the language nuances present in biological texts, hence augmenting the relation extraction framework's overall precision and efficacy.

3.2.3 Syntactic Features:

The structural components of biological sentences are the main emphasis of our methodology's syntactic features. These features help identify verbs connecting named items (drugs and diseases), subject-object pairings, and other syntactic constructions by parsing the syntax, including grammatical relationships and sentence structure. Through a deeper comprehension of the contextual relationships between medications and disorders in biomedical texts, this structural analysis improves the model's capacity to identify grammatical patterns.

3.2.4 Semantic Features:

The method of feature extraction in the given text entails classifying and recognizing words according to their semantic kinds. "TREATment" (TREAT) and "DISease" (DIS) are noteworthy traits. The terms "induce regression" belongs in the TREAT category, indicating a possible treatment,

whereas "metastatic renal cell carcinoma" belongs in the DIS category, indicating a disease. This semantic annotation improves the model's comprehension of the biomedical context of the statement, making it easier to extract significant relationships between certain diseases and possible therapies from the text.

4. Performance Analysis

4.1 *Drug Information Extraction*

Using a machine learning classifier and a classification method with different relation classes CURE, PREVENT, SIDE EFFECT, NO CURE, and OTHER RELATION the links between medications and diseases were extracted. Specific interactions between diseases and therapies were represented by each class. For example, the CURE class represented cases in which medicines successfully reversed disorders, whereas SIDE EFFECT represented diseases brought on by treatments. The algorithm was guided in identifying relationships within biological texts by the categorization, which made use of features extracted in the preceding step.

Owing to the significant amount of classified data, conventional machine learning techniques encountered constraints. Because of its capacity to handle high-dimensional data, a Support Vector Machine (SVM) was selected as the solution. This method improved the model's capacity to categorize complex links between medications and illnesses by effectively separating out individual relationships.

4.2 *Result for Knowledge Discovery in Biomedical*

Table 1 compares evaluation parameters for Knowledge Discovery in Biomedical applications in detail, with three models in particular: Support Vector Machine (SVM), Random Forest (RF), and Decision Tree (DT). Key performance parameters, such as Precision, Recall, F1 Score, Area Under the Curve (AUC), and Accuracy, were used to evaluate these models.

Table 1

Comparison of Evaluation parameter for Knowledge Discovery in Biomedical

Model	Precision	Recall	F1Score	AUC	Accuracy
DT	87.52	86.56	89.66	90.25	86.36
RF	91.25	93.52	92.45	93.12	92.55
SVM	88.65	89.52	89.8	90.35	89.63

SVM demonstrated an excellent precision of 88.65%, while DT trailed closely behind with a respectable 87.52%. In the process of discovering new biological knowledge, precision is essential since it guarantees the validity and applicability of the links found. RF performed exceptionally well, achieving the maximum recall of 93.52%, according to the Recall metric, which measures the capacity to catch all pertinent instances.

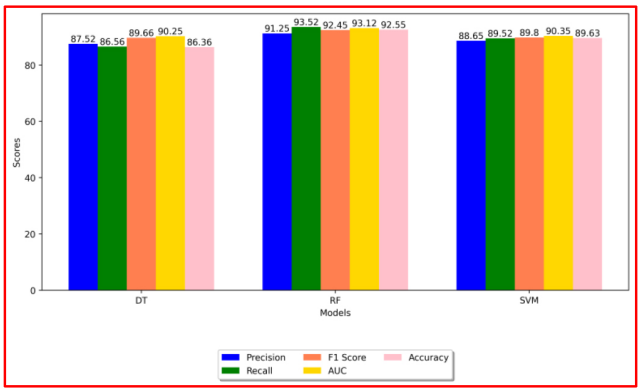

Figure 2

Representation of Evaluation parameter

With a harmonized score of 92.45%, the F1 Score a measure that balances recall and precision exhibited RF as the frontrunner as shown in Figure 2. SVM and DT came in second and third, respectively, with F1 Scores of 89.80% and 89.66%. An essential indicator of the model's capacity to discern between positive and negative occurrences is the Area Under the Curve (AUC).

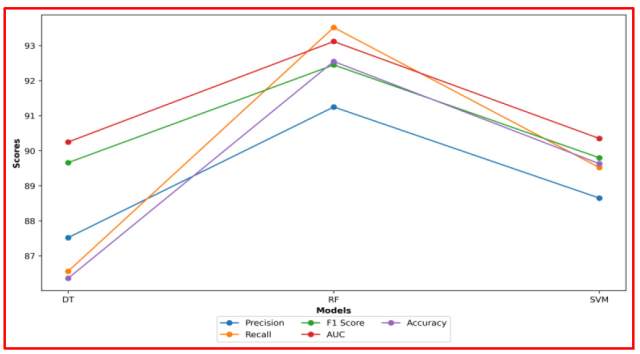

Figure 3

Comparison of Evaluation parameter

With an AUC of 93.12%, RF showed supremacy, highlighting its strong discriminating capacity. Competitive AUC values of 90.35% and 90.25% were shown using SVM and DT, respectively as shown in Figure 3.

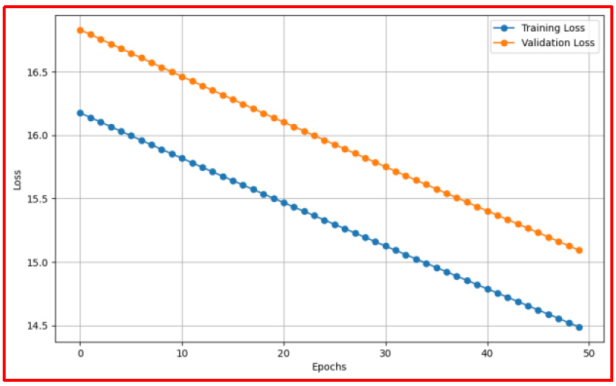

Figure 4

Loss Training and Validation

When it came to accuracy, which measures how accurate all of the predictions were, RF came in first at 92.55%, closely followed by SVM at 89.63% and DT at 86.36% as shown in Figure 4 and 5. This thorough analysis highlights the effectiveness of RF in the discovery of biomedical knowledge, especially when it comes to determining the connections between medications and illnesses.

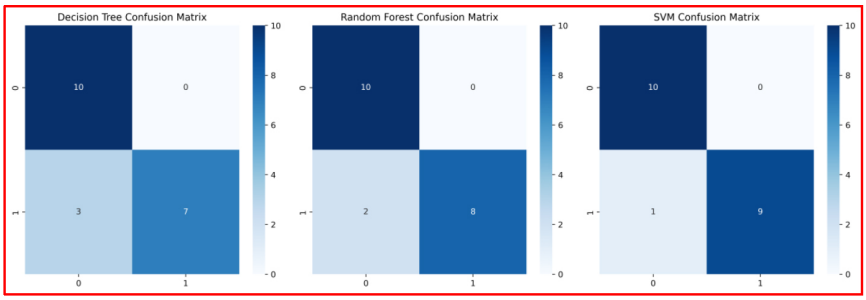

Figure 5

Confusion Matrix

5. Conclusion

Natural language processing (NLP) is being used to extract pharmacological information, which greatly expands our understanding

of what is known about drugs in biological literature. Our all-inclusive strategy, which includes several feature extraction techniques and preprocessing approaches, creates a strong basis for the extraction of drug-disease relations. Finding subtle relationships in biomedical data is improved by integrating machine learning classifiers like Support Vector Machines, Random Forests, and Decision Trees. The comparative evaluation demonstrates that Random Forest performs better in terms of precision, recall, F1 score, AUC, and accuracy, highlighting the effectiveness of ensemble learning in the discovery of biomedical information. This work highlights the value of advanced natural language processing and machine learning for biomedical knowledge discovery, and it makes a substantial contribution to the extraction of pharmacological information. Future developments in drug discovery, clinical decision support systems, and the field of biomedical research as a whole are made possible by the improved approaches that have been described.

References

[1] Y. Zhang and Z. Lu, "Exploring semi-supervised variational auto-encoders for biomedical relation extraction," *Methods,* vol. 166, pp. 112–119 (2019).

[2] Z. Li, Z. Yang, C. Shen, J. Xu, Y. Zhang, and H. Xu, "Integrating shortest dependency path and sentence sequence into a deep learning framework for relation extraction in clinical text," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, p. 22 (2019).

[3] Y. Niu, D. Otasek, and I. Jurisica, "Evaluation of linguistic features useful in extraction of interactions from PubMed; application to annotating known, high-throughput and predicted interactions in I2D," *Bioinformatics,* vol. 26, no. 1, pp. 111–119 (2009).

[4] Y. Zhu, L. Li, H. Lu, A. Zhou, and X. Qin, "Extracting drug-drug interactions from texts with BioBERT and multiple entity-aware attentions," *Journal of Biomedical Informatics,* vol. 106, Article ID 103451 (2020).

[5] T. T. T. Phan, T. Ohkawa, and A. Yamamoto, "Protein-protein interaction extraction from text by selecting linguistic features," in Proceedings of the 2017 IEEE 17th International Conference on Bioinformatics And Bioengineering (BIBE), pp. 181–187, Washington, DC, USA, October (2017).

[6] Y. Niu, H. Wu, and Y. Wang, "Protein-protein interaction identification using a similarity-constrained graph model," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 16, no. 2, pp. 607–616 (2019).

[7] D. Badal Varsha, J. Petras Kundrotas, and A. Ilya Vakser, "Natural language processing in text mining for structural modeling of protein complexes," *BMC Bioinformatics*, vol. 19, no. 1 (2018).

[8] S. Koyabu, T. T. T. Phan, and T. Ohkawa, "Extraction of protein-protein interaction from scientific articles by predicting dominant keywords," *BioMed research international,* vol. 2015, Article ID 928531, 13 pages (2015).

[9] D. Couch, Z. Yu, J. H. Nam et al., "GAIL: An interactive webserver for inference and dynamic visualization of gene-gene associations based on gene ontology guided mining of biomedical literature," *PLoS One*, vol. 14, no. 7 (2019).

[10] Al-Aamri, K. Taha, Y. Al-Hammadi, M. Maalouf, and D. Homouz, "Analyzing a co-occurrence gene-interaction network to identify disease-gene association," *BMC Bioinformatics*, vol. 20, no. 1, p. 70, (2019).