An Empirical Study on Various Word Sense Disambiguation Techniques in the Biomedical Domain

Pawan Makhija[1*] and 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
*[1]pawanmakhija@acropolis.in , [2]sanjay_tanwani@hotmai.com

Abstract. Word sense disambiguation (WSD)in the biomedical domain is a composite task that involves establishing correct meaning of a word based on its explicit context within biomedical literature. Biomedicine literature, constitutes research papers, clinical reports, electronic health records, and pharmaceutical articles. All these sources of biomedical literature are rich with terminology that often has numerous meanings. This leads to uncertainty in drawing an accurate meaning to a word present in the biomedical text. This polysemy introduces a challenge, as the accurate connotation of such terms is important for various subsequent applications that includes information retrieval, text mining, and knowledge extraction. The vagueness in word meanings of biomedical text can lead to misunderstandings or increased errors in biomedical data analysis, thus influencing clinical decision-making, research outcomes, drug discovery and patient's treatment. Therefore, a technique for WSD is required to ensure that the computational systems can precisely process and analyze biomedical literature, that forms a base to more reasonable and fruitful insights. This paper introduces an in-depth investigation of the challenges associated with WSD in the biomedical domain. This investigation includes the structural complexity of medical language, the demand for domainspecific knowledge, and the drawbacks of existing natural language processing (NLP) techniques. We discuss the relevance of WSD in improving the accuracy and efficiency of biomedical data analysis. This study includes several methods, from fundamental rule-based approaches to more advanced machine learning and deep learning models, that are evaluated for their productiveness in addressing WSD challenges in biomedical texts. After conducting an empirical study, the findings showcase effectiveness of BERT model over other machine learning models for the classification problems. Our study also includes exploration of various research that have undergone in this area as well as applications of WSD in biomedical domain.

Keywords: Word Sense Disambiguation(WSD), Natural Language Processing(NLP), Long Short Term Memory (LSTM), Support Vector Machine (SVM), Convolutional Neural Network (CNN)

S. Bhalerao et al. (eds.), *Proceedings of the International Conference on Recent Advancement and Modernization in Sustainable Intelligent Technologies & Applications (RAMSITA-2025),*Advances in Intelligent Systems Research 192,
https://doi.org/10.2991/978-94-6463-716-8_79

1 INTRODUCTION

In the ever-expanding biomedical field, there is an ample amount of text data coming from various sources like biomedical research, doctor's prescriptions, patient reports, articles, research papers, drug discovery, EHR etc that presents a challenge for natural language understanding systems. This challenge presents a critical s hurdle in the ambiguity associated with many medical terms, thereby entailing research in the word sense disambiguation applications techniques (Fig. 1). WSD aims to identify the correct meaning of words within the specified context, thus ensuring the actual interpretation of the medical term and strengthening the effectiveness of various applications in healthcare informatics. Word Sense Disambiguation in the biomedicine is an auxiliary task of Natural Language Processing. It deals with the difficulty of identifying the correct meaning of a word in a particular context. For an example from the clinical texts coming various sources, the term "blood pressure" has three possible meanings. Now according to the Unified Medical Language, its having a sense of organism function, diagnostic procedure, and laboratory test result [1]. There are certain cases where many words in natural language have multiple different meanings. Here WSD aims to relate the correct sense of a word in a given particular context. With the augmented growth in clinical literature from various sources and the complexity of its terminology, ensuring an exact understanding of word meaning is important for information extraction, knowledge extraction, text mining, information retrieval, and decision-making. A correct sense or meaning of the clinical text facilitates the extraction of meaningful information that, in turn, can lead to exponential progress in the areas like drug discovery, disease diagnosis, medical research, medicine repurposing and personalized medicines.

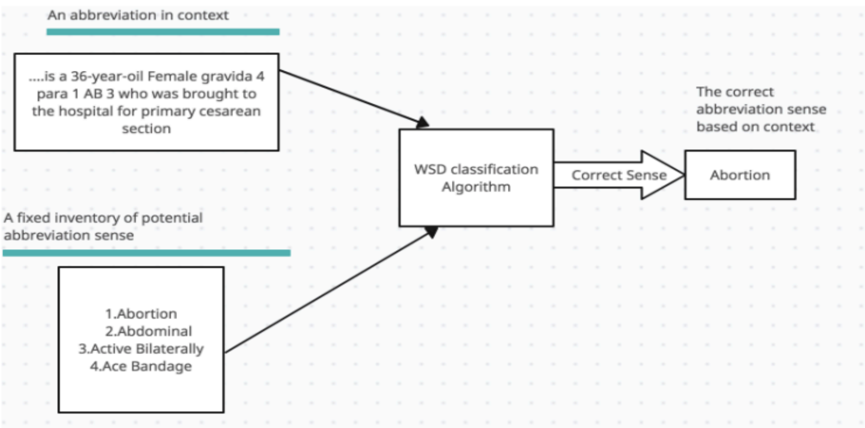

Fig. 1. Word Sense Disambiguation process for medical abbreviations

Strategies for Word Sense Disambiguation in biomedical domain comprises of:

(i) Supervised Machine Learning: This type of learning makes use of labelled datasets for training model that can predict the target values based on previous instances. In case of Supervised machine learning, the training of model is done on a dataset containing annotated data. Here each of the sample in biomedical field is mapped with the correct sense. In supervised learning, each data sample consists a target word and is associated with correct meaning in a specific context. In the later stages, the model helps to forecast the correct meaning of the target word in specific contexts. Supervised machine learning algorithms used in the biomedical domain encompasses support vector machine (SVM), decision trees, naïve bias, random forest and neural networks.

(ii) Unsupervised Machine Learning: It includes grouping words that share the same contexts and then allocating their correct meaning to the resultant clusters. This method does not mandate annotated biomedical text data for training. This in turn tends to be less precise as compared to supervised learning methods. This learning used in biomedical data includes word embeddings (word2Vec, GloVe), clustering algorithms (k-means, hierarchical clustering), graph-based algorithms (PageRank, label propagation) etc.

(iii) Knowledge Based approaches: It requires the use of explicit knowledge base, like a dictionary or ontology, to differentiate semantic relationships between the words and their meanings associated in biomedical text. Basically, the effectiveness of this approach is based on the quality and accessibility of the knowledge base available in the data. Commonly used knowledge-based approaches are Wordnet based, ontology based and Lesk algorithm that has their leveraging resources like UMLS or SNOMED CT.

(iv) Hybrid Methods: This mainly incorporates the different approaches altogether, such as combining supervised and knowledge-based techniquesin a single model to achieve better accuracy. It basically focuses to enhance the strengths of algorithm by improving the precision of the WSD process.

2 RELATED WORK:

Al-Mubaid H, Gungu S proposed a supervised learning approach for word sense disambiguation in medical domain. This methodology used a NLM-WSD dataset, that enabled a direct improvement with findings from previous studies. For building the model, they used support vector machine that achieved mean accuracy around 90.3%. In the later stages to validate their research work, 5-cross validation algorithm was used.[1] Pesaranghader et al. introduced the deepBioWSD model that uses a bidirectional Long Short-Term Memory (BiLSTM) network to disambiguate any medical term with various meanings in biomedical literature. Their research work used the MSH WSD dataset. The model achieved an accuracy of about 96.82%.[2] Zhang C, Biś D, Liu X, He Z. presented paper that is based on two deep-learning-based models. Their research work was designed for supervised Word Sense Disambiguation in the biomedical textual data. The two techniques used in this paper are a Bi-directional Long Short-Term Memory (BiLSTM) network model and an attention model that utilizes self-attention architecture. This research study used MSH WSD dataset for model building. Among the two techniques the BiLSTM neural network model, constitutes appropriate upper layer configuration, that increased its performance from the existing models. In comparison to BiLSTM model, the attention-based model was used in this paper. This lead to the further enhancement of the accuracy levels, thereby highlighting the effectiveness of attention mechanisms in WSD tasks in medical domain. This research introduced model that disambiguate the ambiguous words by combining the embedding of the target ambiguous word with the maxpooled vector. This combined technique resulted in an accuracy rate of nearly 90 % for the WSD in biomedical domain. [3] Skreta M, Arbabi A, Wang J, Drysdale E, Kelly J, Singh D, Brudno M presented a research paper that focused on improving word sense disambiguation in biomedical text through the implementation of data augmentation. This method makes use of information processing from biomedical ontologies. They trained their model by using convolutional neural network (CNN) on the openly available datasets. They integrated word embeddings with the ontologies on MIMIC-III corpus. Results proved that the incorporation of ontology-informed deep learning techniques have broader implications for varied biomedical activities beyond abbreviation disambiguation. The capacity of this approach drives the advancements in machine learning within healthcare informatics.[4] Kim S, Yoon J presented paper that introduces an approach combining convolutional neural networks (CNN) and support vector machines(SVM) for word sense disambiguation within the biomedical domain. Their research used three distinct datasets to achieve objective of WSD in biomedical text data. The incorporation of convolutional neural network model has resulted in a notable improvement in accuracy of the model. In the later stages an indepth model validation was carried out for both convolutional neural networks and support vector machines through 5-fold cross-validation and 10-fold cross-validation techniques.[5] Moon S, Berster BT, Xu H, Cohen T introduced a Binary Spatter Code Word Sense Disambiguation (BSC-WSD) algorithm that uses hyper dimensional computing. It resulted into a notable level of accuracy for resolving ambiguities in the biomedical data and achieved the accuracy of about 94.55% by the one-to-one BSC-WSD method. They utilized hyper dimensional computing to perform vector transformations. These encode ambiguous terms and maps with their context-specific meanings.[6] Berster BT, Goodwin JC, Cohen T, presented a methodology for the Word Sense Disambiguation in biomedical domain. It introduced, use of binary vectors for the implementation of a Vector Symbolic Architecture (VSA). The application of this methodology relates to the issues like acronym disambiguation, and the continued evaluation of distributed representations of similar nature as a mechanism for executing classification tasks across various biomedical domains.[7] Jane Millar, presented a paper highlighting the importance and benefits of SNOMED

CT in healthcare. SNOMED CT is a crucial asset in the field of healthcare informatics. Their research idea in the field of WSD in biomedical domain provided a comprehensive and multilingual medical terminology on a globally large scale. This fusion of system guarantees a uniform and dependable depiction of clinical information in electronic health records (EHRs), offering advantages to individual patients, healthcare providers, and populations. Through the advancements of evidence-based practices and the improvement of communication through Electronic Health Records, SNOMED CT assumes an important role in enhancing the quality of healthcare informatics [8]. Study of Literature Survey is given in Table 1.

Table 1. Study of Literature Survey

S.no	Title	Technique	Year
1	A learning-based approach for biomedical word sense disambiguation.[1]	SVM and 5-fold cross-validation	2012
2	A deepBioWSD: effective deep neural word sense disambiguation of biomedical text data.[2]	Bidirectional Long short-term memory	2019
3	Biomedical word sense disambiguation with bidirectional long short-term memory and attention-based neural networks.[3]	Bidirectional Long short-term memory and attention-based neural networks	2019
4	Automatically disambiguating medical acronyms with ontology-aware deep learning.[4]	CNN	2021
5	Link-topic model for biomedical abbreviation disambiguation. [5]	CNN,5-fold cross-validation and 10-fold cross-validation	2015
6	Word Sense Disambiguation of clinical abbreviations with hyperdimensional computing.[6]	BERT model	2013
7	Hyperdimensional computing approach to word sense disambiguation. [7]	RNN	2012
8	The Need for a Global Language - SNOMED CT Introduction. [8]	BERT model	2016
9	Natural Language Processing to extract SNOMED-CT codes from pathological reports. Pathologica. [9]	NLP	2023
10	BertSRC: transformer-based semantic relation classification.[10]	BERT model	2022
11	Semantic annotation in biomedicine: the current landscape.[11]	Word Embeddings	2017

12	Biomedical Word Sense Disambiguation with Word Embeddings. [12]	Word Embeddings	2017
13	MeDAL: Medical Abbreviation Disambiguation Dataset for Natural Language Understanding Pretraining. [13]	NLP	2020
14	Clinical Abbreviation Disambiguation Using Neural Word Embeddings.[14]	Word Embeddings	2015
15	Context Embedding Based on Bi-LSTM in Semi-Supervised Biomedical Word Sense Disambiguation.[15]	Bidirectional Long short-term memory	2019
16	Knowledge-Based Biomedical Word Sense Disambiguation with Neural Concept Embeddings.[16]	Knowledge Based	2017
17	Supervised Learning and Knowledge-Based Approaches Applied to Biomedical Word Sense Disambiguation.[17]	Supervised machine learning techniques	2017
18	Applying active learning to supervised word sense disambiguation in MEDLINE. [18]	Supervised learning	2013
19	Interactive medical word sense disambiguation through informed learning.[19]	Informed Learning	2018
20	Disambiguation in the biomedical domain: the role of ambiguity type. [20]	Supervised learning	2010
21	Determining the difficulty of Word Sense Disambiguation.[21]	Word embeddings	2014
22	Graph-based Word Sense Disambiguation of biomedical documents.[22]	NLP	2010
23	Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification.[23]	CNN	2013
24	Word embeddings and recurrent neural networks based on Long-Short Term Memory nodes in supervised biomedical word sense disambiguation. [24]	Wordembeddings, RNN and Bidirectional Long short-term memory	2017

3 PROPOSED METHODOLOGY:

3.1 Dataset Description:

This study used two primary datasets: First one is the annotated abbreviation dataset that is sourced from the University of Minnesota (UMN) National Institute of Healthcare (NIH). Second dataset is the medal dataset that is obtained from Kaggle. After the acquisition, the data was converted into excel spreadsheet format from its original text. The dataset sums up the information, with 75 unique abbreviations, within 500 varied contexts. The dataset contains three columns—abbreviations, sense, and context with a total of 37,500 rows in number. This dataset serves the purpose for our research in biomedical abbreviation disambiguation by classification (Fig. 2).

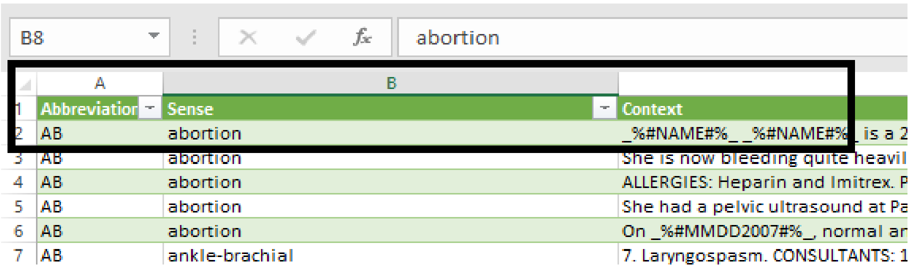

Fig. 2. Sample Dataset for disambiguation process

3.2 Data Pre-processing

In our analysis, we implemented three supervised machine learning algorithms, each of which has contributed efficacy in Word Sense Disambiguation. Vector Space Model (VSM) is a memory-based learning that is an algebraic model. It represents text documents in the form of a N-dimensional vector and this N denotes the number of distinct terms in the available text. It represents target word in the form of a binary vector, where each position shows the presence or absence of a feature. In the training centroids are used to denote each viewpoint, and similarity with new models is constituted using the cosine similarity metric. Then in the later stages nearest centroid distance designates the point assigned to the new model, thus helps in finding correct meaning of a word.

Conceptually in the Vector Space model, initial coefficient relates to the distance of surrounding words from the target word within the textual corpus. Then this distance measurement is characterized not only by the closeness in the sentence, but rather by the estimation of the Euclidean distance of target word within the vector space in relation to each of the nearest adjacent words in its context.

The rationale behind the utilization of this coefficient lies in the premise that the words that are exhibiting closeness in terms of vector space distance, indicating semantic similarity, are expected to carry a more prominent weight as compared to words that are placed at more distance that ultimately lacks semantic similarity. A mathematical equation is presented to incorporate word vector similarities that sets with the above-mentioned principles is as follows:

$$E_i = w_{ij} \ (1 - \alpha)^{|I-j|-1} \quad (1)$$

The main score functions that are used in our study are:
1. Term Frequency(tf)
2. Inverse term frequency(tf-idf)

Term Frequency: In a target document term frequency represents the occurrence of a particular word. These terms are assigned the weights that are proportional to the frequency of a term and is denoted by (tf).

For example, given word (t) and a document (d) is represented by d then weight is computed as:

$$\text{Tf}(t, d) = (\text{count of } t \text{ in } d) \ / \ (\text{total of words in } d) \quad (2)$$

Inverse Document Frequency: It shows the importance of word over a entire corpus and focuses to tests how relevant the word is in that particular document.

Tf-idf score: This combines both term frequency and inverse term frequency and shows the importance of a term in a document. High values of tf-idf shows the relevancy of a word in the specified document.

3.3 Model Design through various Techniques:

Next stage is the model building which deals with the selection of machine learning model for the task of classification. This paper comprises of a comparative study of implementation of various supervised machine learning models for WSD in biomedical domain.

First approach in our study is logistic regression. Basically, logistic regression is a statistical modelling that is focused to predict the probability of a instance belongs to a given class or not based on input features. Logistic regression, is applied to the WSD.xlsx dataset. This has showcased the efficacy of this method in identifying the correct word senses in the biomedical domain from the large amount of literature.

Secondly, the Naive Bayes (NB) classifier was implemented that uses the concept of Bayes' Theorem for training the model and operates on a probabilistic model. It has the assumption that the classification is independent of given features present in the dataset. It uses the concept of posterior probability and considers that one of the features in our model does not depend on the existence of another feature in the model. This probabilistic approach finds a solution for word sense disambiguation in biomedical domain by calculating the likelihood of correct sense in the particular context.

Thirdly Support Vector Machine (SVM), is another conventional classification technique used in our study. SVMs has a capability to examine any data and understand patterns, this can help to classify input into different available categories. Due to this robustness in SVMs it can be used for numerous tasks like text or image classification, spam filtering, biometrics, anomaly detection. It is based on the fact that maps feature vectors to a high-dimensional space and can even manage the non-linear relationships. It then contributes a classifier by identifying the plane that maximally separates the classes. SVM's are beneficial as it can handle high-dimensional data and it also has the ability to discover complicated decision boundaries thus making it a good choice for WSD in clinical domain. The implementation with SVMs was done on the WSD.xlsx dataset. Thereby contributing in resolving ambiguity in determining the correct meaning of any word in biomedical domain.

Next model is the implementation of Decision Tree algorithm on WSD.xlsx dataset. This algorithm is based on the fact that it creates the decision trees that recursively partitions the data that is based on the various features. With this level of accuracy it shows that it is capable of capturing patterns that are important for word sense disambiguation in the specialized biomedical domain. The findings show represents its capability in classifying various contexts in biomedical literature, thus contributing to the overall success of the WSD process.

Random forest is another model that was implemented on the WSD.xlsx dataset. It is the ensemble learning that constructs a large number of decision trees that can enhance the predictive modelling. Due to this power of random forest, it has found its application for the task of WSD in the biomedical domain.

3.4 Evaluation Metrics:

Evaluation metrics forms a basis for quantitative analysis that is used to measure the performance of a machine learning model through its different metrics. After the model building another thing comes into play is the model evaluation with the help of various evaluation metrics that can measure the performance and effectiveness of our model. Our study enlists the following evaluation metrics in detail.

Classification Accuracy (ACC) score.

Classification Accuracy score is the ratio of correctly predicted instances to the total of all the possible instances.Accuracy score gives a knowledge of how well a classification model performs in the terms of correct predictions. In case of biomedical domain class distributions sometimes may be imbalanced, accuracy alone might not be sufficient to evaluate the model’s performance. Mathematically accuracy score can be calculated as:

```
Accuracy = (No. of correct predictions)/ (Total no. of input samples)
```
(3)

Precision.

Precision is another evaluation metric that guides us that how many positive predictions are made by the model are actually correct. It is playing a significant role when the cost of false positive values are high. In the case of biomedical text data correct sense of word is of utmost important as incorrect sense can lead to wrong interpretations or diagnoses that in turn can end with a wrong treatment to a particular diagnosis. Precision plays a critical role, especially in clinical applications where misinterpretations of a particular term or misdiagnoses of a disease can have serious implications on a person's health. Mathematically precision is the ratio of positive correct values in the prediction to the total number of positive predicted values.

```
Precision = (True positive values)/(True positives) + (False
Positives)
(4)
```

Recall or Sensitivity.

Recall is the evaluation metric quite similar to precision. Basically, it is ratio of true positive predictions that are actually true to that of total number of positive values that can be both correctly predicted as positive values or correctly predicted as false values. Now in case of WSD in biomedical domain one need to ensure that most of the correct senses are identified. Recall denotes the sensitivity of a classifier. As WSD is a kind of application in which identification of correct meaning to a word in context can have critical implications that cannot be over looked. Mathematically recall or sensitivity can be denoted by the below stated formulae:

```
Recall = (True positive values)/(True positives) + (False Negatives)
```
(5)

F1-Score.

Basically F1-score is the evaluation metric that can be calculated as the harmonic mean of both Precision and Recall. F1-score gives a single evaluation metric that balances the trade-off between both precision and recall. It is usual to use F1- score metric when the distribution of class is imbalanced. In case of WSD in biomedical domain evaluation of a model can be done on the basis of F1- score as it is a combination of both precision and recall so, proper sense of a word in a biomedical context will be based on sensitivity. Mathematically F1- score can be represented by the formulae stated below:

```
F1-Score=2x [(Precision x Recall)/(Precision + Recall)]
```
(6)

Specificity.

It is the ratio of correctly predicted negative observations to the total actual negatives. Higher value of specificity indicates that the model can correctly tell the information about most of its negative results. So when specificity is used as a evaluation metric in biomedical domain then it can tell the information where the model can predict the false values. Specificity is important in the contexts where it's important to correctly identify the absence of correct senses, thereby avoiding the false positives. Mathematically specificity can be denoted as:

```
Specificity=(True Negatives)/(False Positives+(True Negatives)
```
(7)

Confusion Matrix.

A confusion matrix is a evaluation metrics consisting of a table. It is used to measure the performance of a model on a set of test data. It basically displays the number of instances that are accurate or not based on correct predictions that are made on the model. It provides detailed description of model's performance by illustrating how many instances were correctly or incorrectly classified for each class. In the case biomedical domain, it can help us to identify a correct sense made to a word in a context is correct or not. These values can then be labelled as under the following headings in the table shown below.

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Explanation of all the instances of confusion matrix showing the number of instances produced by the model on the test data as under these headings.

1.True Positive (TP): Model correctly predicted a positive outcome and the actual outcome was also positive.

2.True Negative (TN): Model correctly predicted a negative outcome and the actual outcome was negative.

3.False Positive (FP): Model incorrectly predicted a positive outcome and the actual outcome was negative.This is known as a Type I error.

4.False Negative (FN): Model incorrectly predicted a negative outcome and the actual outcome was positive.This is known as a Type II error.

4 CHALLENGES OF WSD IN BIOMEDICAL DOMAIN:

Word sense disambiguation (WSD) finds various challenges in biomedical domain due to the its complexity associated with its terminology. In the biomedicine informatics specificity of its terminology, is of utmost importance otherwise if this not taken proper care can lead to unfortunate consequences on once health. This in turn end up with the complexities that are inherent for the natural language. Some of the main challenges are listed below:

4.1 **Polysemy and Homonymy:**

In biomedical literature, polysemy and homonymy are frequently seen. Polysemy means that a single term in biomedical domain may have multiple meanings. On the other side homonymy deals with the fact that a single terminology in biomedical domain may have multiple unrelated meanings. This presents a challenge for the researchers in medical field to resolve these types of ambiguities.

For Example:

- Cell: It may have a meaning as biological cell, a unit of measurement, or a social group.
- Delta: Similarly, in the biomedical literature, it can refer to the Delta variant of a virus or delta can be used in mathematical contexts too.

It becomes necessary to resolve these ambiguities present in the biomedical literature by considering its contextual usage that is grounded in domain knowledge expertise.

4.2 **Specialized Terminology:**

The domain of biomedical literature constitutes huge vocabulary that is specific to different disciplines like genetics, biochemistry, microbiology, drug discovery and pharmacology. Various unusual terms within this biomedical domain may lack sufficient representation of knowledge in natural language resources. This in turn complicates the development of precise WSD systems in biomedical domain in the absence of predefined domain knowledge.

4.3 **Data Sparsity:**

The annotated datasets that are dedicated to WSD in biomedical literature are very much constrained in terms of size and inclusivity. The foundation of biomedical datasets may be manually assigned meanings for a broad range of its terms is resource-intensive and time-consuming. This in turn, WSD models may find challenges in generalizing across a vast contexts and terms. Thus, leading to low performance, especially concerning difficult or specialized terms.

4.4 **Ambiguity Resolution:**

Many a times the WSD task in biomedical terms demands consideration not only of the immediate context but also of the broader semantic and syntactic context within a document or corpus. Now addressing such a ambiguity at the individual term level that is accounting for the correct sense of a sentence represents a difficult job thus necessitating refined linguistic and domain-specific intelligence.

4.5 Ambiguity in Abbreviations and Acronyms:

Basically, the biomedical literature is rich with large number of abbreviations and acronyms, many of which can lead to ambiguity and can encompass multiple meanings depending on the particular context. The accurate and efficient interpretation of biomedical text focuses on resolving the correct sense of abbreviations and acronyms that is quite a complicated task as there is the absence of proper definitions and a knowledge base.

4.6 Semantic Drift and Evolution:

The semantic support of biomedical terms are subjected to the evolution over time with newer terms in biomedical domain. This may lead to the evolution of newer scientific research. WSD task in biomedical domain must be adaptable to these transformations, thus identifying the new emerging senses of terms with greater precision and proper consistency.

Exploring these above defined challenges mandates merging of domain-specific knowledge, systematic NLP methodologies, and innovative ideas for harnessing large amount of available data and resources.

5 APPLICATIONS OF WORD SENSE DISAMBIGUITION IN BIOMEDICAL DOMAIN

Word Sense disambiguation plays significant role in various applications in the biomedical domain that in turn helping health care related tasks and systems. There are many progressive and specialized uses of word sense disambiguation (WSD) in the biomedical domain. This can influence evolution of cutting-edge technologies and methodologies. These significant applications of WSD in the clinical domain explains the strength of this technology that finds a solution to such complex challenges in the field of healthcare. It is contributing to recent advancements in biomedical research, patient care, improved healthcare systems, drug discovery and public health.

Some of the applications of WSD in biomedical domain are listed below:

5.1 Clinical Ontology Mapping:

WSD techniques are applied to map clinical terms to basic medical ontologies such as SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) or UMLS

(Unified Medical Language System) [9][10]. Its capability for identification of correct sense for a given clinical term from various sources like electronic health records, has contributed to achieve a milestone for the healthcare providers in many different ways.[8] Disambiguating the biomedical terms and mapping them to a correct ontology concept, WSD has enhanced the interoperability between different healthcare systems and formed a base for more efficient data integration, retrieval and analysis.

5.2 Semantic Annotation of Clinical Text:

WSD in biomedical domain can be used to semantically to describe clinical data with standardized and generalized medical concepts and draw relationships between those concepts. By disambiguating the biomedical terms and properly linking them to the ontology concepts, clinical documents can be enriched with structured metadata. Thus, WSD task is helping for more better information retrieval, medical data analysis, and provides the decision support for drug discovery.[11][12]

5.3 Clinical Phenotyping and Biomarker Identification:

Many a times there are multiple meanings associated with clinical phenotypes or biomarkers. Basically, clinical phenotype deals with the characteristics of a particular disease in an individual. So, WSD techniques in biomedical domain can help in describing these phenotypes. Biomarkers explains the molecule or gene associated with a pathological or physiological impacts of a disease. Task of WSD bridges a gap by disambiguating the ambiguous terms associated with a disease, its symptoms, and biomarkers. By the accurate interpretation of clinical text, WSD can help information extraction related to phenotypes that are coming from electronic health records and biomedical literature, research and personalized medicine initiatives.

5.4 Adverse Drug Event Detection:

A drug can have positive as well as a adverse impact on an individual's health. WSD in biomedical domain can play an important role in the detection of adverse effects of a drug in many events (ADEs) that are originating from various sources. WSD has also found its application in disambiguating the drug-related terms and differentiating between its different types of drug reactions and interactions. By accurate interpretations of drugs and adverse events in clinical text, WSD helps improve the performance of Adverse drug events (ADE) detection algorithms and enhances pharmacovigilance efforts.

5.5 Clinical Concept Normalization:

The task of WSD in biomedical domain can be applied to normalize biomedical concepts mentioned in text to a standard representation. As an example, it can map a medical ontology concept or a controlled vocabulary term with its proper meaning. By this means disambiguating the clinical terms, WSD contributes for proper data integration, aggregation, and its analysis in different sources. Thus, enabling more advanced clinical research and decision-making.

5.6 Multimodal Clinical Data Integration:

WSD can extend its application to integrate information from multiple modalities that comprises of text, images, and sensor data, in clinical data to enhance the performance of decision support systems and diagnostic tools. By disambiguating medical terms and mapping them with other biomedical data sources, WSD enables a better understanding of an individual patient conditions and contributes in more accurate diagnosis and treatment of a disease.

6 Results

We have conducted the empirical study on various machine learning models especially for biomedical word sense disambiguation. The purpose of this research is to identify the best models for the disambiguation process. The results shown that, BERT model outperformed in addressing the challenges possessed by the ambiguous text.Various performance metrics of machine learning has been evaluated. The comparative study of various algorithms are as follows given in Table 2:

Table 2. Performance metrics of various models experimented for WSD

S.no	Technique	Accuracy	Precision	Recall	F1-Score
1	Logistic Regression	84.14%	0.82	0.84	0.78
2	Support Vector Machine	84.86%	0.84	0.85	0.83
3	KNN	62%	-	-	-
4	Decision Trees	85.5%	0.85	0.85	0.85
5	Random Forest	88.43%	0.88	0.89	0.88
6	Bert Model	91.2%	0.90	0.91	0.89

6.1 **Explanation:**

The bar graph illustrates the accuracy (ACC) of different machine learning models used for the word sense disambiguation (WSD) task in biomedical data. The models that we have taken into consideration are Logistic Regression (LR), Support Vector Classifier (SVC), K-Nearest Neighbours (KNN), Decision Tree (DT), Random Forest (RF) and BERT model.

According to the implementation of all algorithms the findings suggest that conventional linear classifiers like logistic regression and support vector machine when combined with tree-based approaches like decision trees and random forest exhibits good performance for WSD in biomedical domain. Random Forest model proved itself good in terms of performance at a accuracy level of 88.5% (shown in Fig. 3).

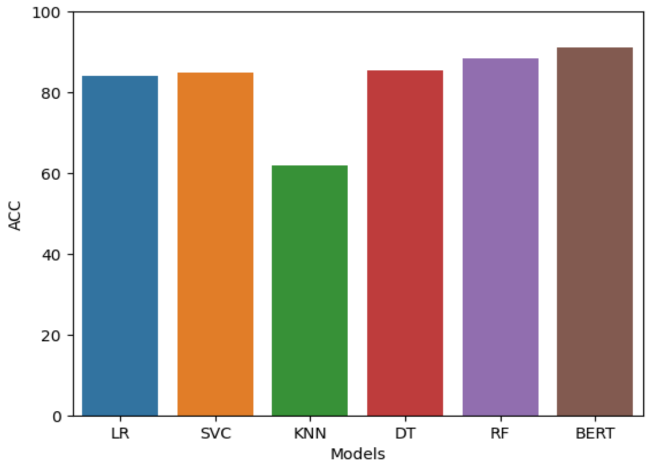

Fig. 3. Comparison of accuracy of various classification models

On the other hand, the performance of the KNN model does not work well in terms of performance in the context of WSD in biomedical domain. The low accuracy achieved with KNN shows its dependency on the concept of distance. This shows that the efficacy of KNN is very much influenced by the selection of value of 'k'. This in turn, necessitates fine-tuning efforts to get good results. In later stages of implementation, the BERT model demonstrated a higher level of accuracy that is around 92%.

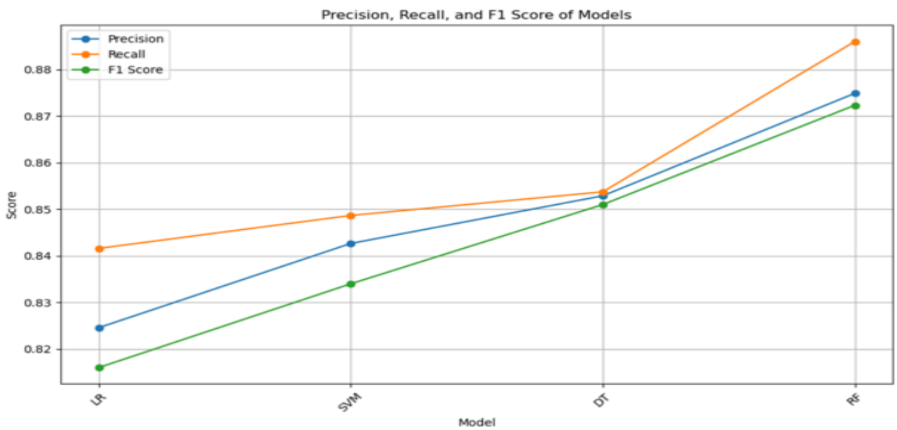

Fig. 4. Evaluation metrics score of various classification algorithms

The line chart (Fig. 4) represents a wide range of evaluation metrics such as precision, recall, and f1 score corresponding to various machine learning models that have been implemented in the word sense disambiguation (WSD) in the biomedical domain. The machine learning models that have been fitted on WSD.xslx data includes the Logistic Regression (LR), Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), Decision Tree (DT), and Random Forest (RF). In the later stages of model development BERT model is implemented that has given the highest accuracy.

The experimentation results showcase that in the task of WSD in biomedical domain, the conventional linear classification approaches like LR and SVC when implemented with tree-based methodologies (DT and RF) exhibits good performance.

From the results it can be inferred that Decision trees has given the outstanding performance with the precision level of approximately 0.86. Subsequently, the Random Forest model demonstrated the bit higher precision of around 0.88 on the same dataset.

It can be seen from the visual representation that the highest recall value is attained through the tree-based approaches like decision trees and random forest. Random forest confirms the recall rate of 0.89, whereas it is a slight lower value, for recall in decision trees.

Similarly, F1 score attained by the random forest is 0.87, that stand more in decision trees by 2 points.

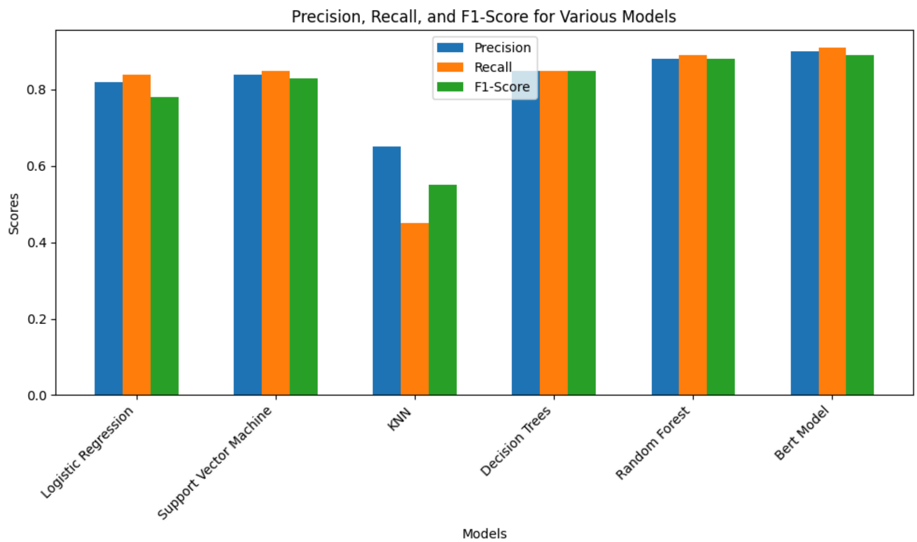

Fig. 5. Precision, recall and F1- score of various machine learning models

Here bar graph (Fig. 5) shows a complete visualization of the three scores namely precision, recall and F1 score for all the models that are used in our study. This bar graph is a visualization of evaluation metrics for various machine learning approaches accompanied with the BERT model in our comparative analysis of WSD in biomedical domain.

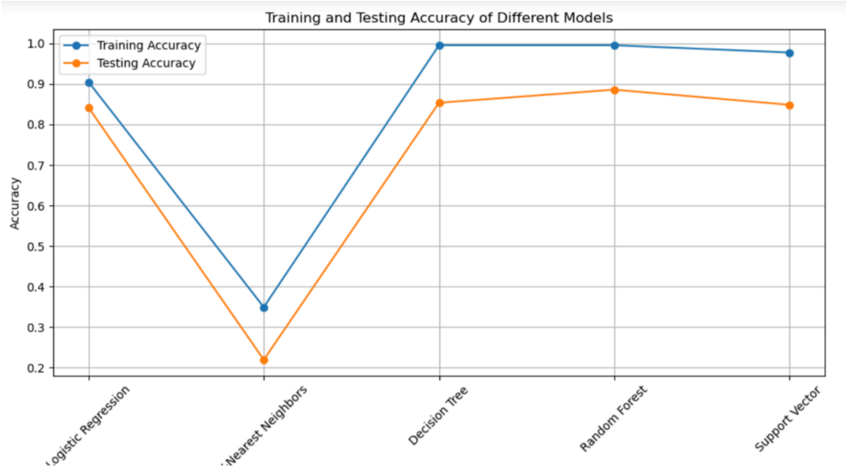

Fig. 6. Training and testing accuracy of various models

In our comparative study of WSD in biomedical domain training and testing accuracies for different models is shown in Fig. 6. From this it becomes clear that random forest and support vector machine has given good results thereby achieving the higher accuracies.

On the other hand, the K-Nearest Neighbors model falls short in accuracy score. In case of, the Decision Tree model, it is giving high accuracy during the training phase but then it suffers overfitting issues that is ultimately leading to low performance in testing phase.

7 CONCLUSION

This paper showcases the empirical results put forth for task of WSD in biomedical domain. It also provides an idea about KNN that is having low accuracy in this domain of biomedicine. There existother alternative models, in our study including Logistic Regression (LR), Support Vector Classifier (SVC), Decision Trees (DT), and Random Forest (RF), that represents higher levels of accuracy. The BERT model used in our study emerged as the most accurate among other algorithms. BERT model is found to be more accurate and proficient in terms of accuracy, precision, recall in WSD task in biomedical domain.

REFERENCES

1. Al-Mubaid H, Gungu S. A learning-based approach for biomedical word sense disambiguation. ScientificWorldJournal. 2012;2012:949247. doi: 10.1100/2012/949247. Epub 2012 May 1. PMID: 22666174; PMCID: PMC3361294.
2. Pesaranghader A, Matwin S, Sokolova M, Pesaranghader A. deepBioWSD: effective deep neural word sense disambiguation of biomedical text data. J Am Med Inform Assoc. 2019 May 1;26(5):438-446. doi: 10.1093/jamia/ocy189. PMID: 30811548; PMCID: PMC7787358.
3. Zhang C, Biś D, Liu X, He Z. Biomedical word sense disambiguation with bidirectional long short-term memory and attention-based neural networks. BMC Bioinformatics. 2019 Dec 2;20(Suppl 16):502. doi: 10.1186/s12859-019-3079-8. PMID: 31787096; PMCID: PMC6886160.
4. Skreta M, Arbabi A, Wang J, Drysdale E, Kelly J, Singh D, Brudno M. Automatically disambiguating medical acronyms with ontology-aware deep learning. Nat Commun. 2021 Sep 7;12(1):5319. doi: 10.1038/s41467-021-25578-4. PMID: 34493718; PMCID: PMC8423722.
5. Kim S, Yoon J. Link-topic model for biomedical abbreviation disambiguation. J Biomed Inform. 2015 Feb;53:367-80. doi: 10.1016/j.jbi.2014.12.013. Epub 2014 Dec 30. PMID: 25554684.
6. Moon S, Berster BT, Xu H, Cohen T. Word Sense Disambiguation of clinical abbreviations with hyperdimensional computing. AMIA Annu Symp Proc. 2013 Nov 16;2013:1007-16. PMID: 24551390; PMCID: PMC3900125.
7. Berster BT, Goodwin JC, Cohen T. Hyperdimensional computing approach to word sense disambiguation. AMIA Annu Symp Proc. 2012;2012:1129-38. Epub 2012 Nov 3. PMID: 23304389; PMCID: PMC3540565.
8. Millar J. The Need for a Global Language - SNOMED CT Introduction. Stud Health Technol Inform. 2016;225:683-5. PMID: 27332304.
9. Cazzaniga G, Eccher A, Munari E, Marletta S, Bonoldi E, Della Mea V, Cadei M, Sbaraglia M, Guerriero A, Dei Tos AP, Pagni F, L'Imperio V. Natural Language Processing to extract SNOMED-CT codes from pathological reports. Pathologica. 2023 Dec;115(6):318-324. doi: 10.32074/1591-951X-952. PMID: 38180139; PMCID: PMC10767798.
10. Lee Y, Son J, Song M. BertSRC: transformer-based semantic relation classification. BMC Med Inform Decis Mak. 2022 Sep 6;22(1):234. doi: 10.1186/s12911-022-01977-5. PMID: 36068535; PMCID: PMC9446816.
11. Jovanović J, Bagheri E. Semantic annotation in biomedicine: the current landscape. J Biomed Semantics. 2017 Sep 22;8(1):44. doi: 10.1186/s13326-017-0153-x. PMID: 28938912; PMCID: PMC5610427.
12. Rui, Antunes., Sérgio, Matos. (2017). Biomedical Word Sense Disambiguation with Word Embeddings. doi: 10.1007/978-3-319-60816-7_33
13. Zhi, Wen., Xing, Han, Lu., Siva, Reddy. (2020). MeDAL: Medical Abbreviation Disambiguation Dataset for Natural Language Understanding Pretraining. 130-135. doi: 10.18653/V1/2020.CLINICALNLP-1.15
14. Yonghui, Wu., Jun, Xu., Yaoyun, Zhang., Hua, Xu. (2015). Clinical Abbreviation Disambiguation Using Neural Word Embeddings. 171-176. doi: 10.18653/V1/W15-3822
15. Zhi, Li., Fan, Yang., Yaoru, Luo. (2019). Context Embedding Based on Bi-LSTM in Semi-Supervised Biomedical Word Sense Disambiguation. IEEE Access, 7:72928-72935. doi: 10.1109/ACCESS.2019.2912584

16. A.K.M., Sabbir., Antonio, Jimeno-Yepes., Ramakanth, Kavuluru. (2017). Knowledge-Based Biomedical Word Sense Disambiguation with Neural Concept Embeddings. 2017:163-170. doi: 10.1109/BIBE.2017.00-61
17. Antunes R, Matos S. Supervised Learning and Knowledge-Based Approaches Applied to Biomedical Word Sense Disambiguation. J Integr Bioinform. 2017 Dec 13;14(4):20170051. doi: 10.1515/jib-2017-0051. PMID: 29236676; PMCID: PMC6042812.
18. Chen Y, Cao H, Mei Q, Zheng K, Xu H. Applying active learning to supervised word sense disambiguation in MEDLINE. J Am Med Inform Assoc. 2013 Sep-Oct;20(5):1001-6. doi: 10.1136/amiajnl-2012-001244. Epub 2013 Jan 30. PMID: 23364851; PMCID: PMC3756255.
19. Wang Y, Zheng K, Xu H, Mei Q. Interactive medical word sense disambiguation through informed learning. J Am Med Inform Assoc. 2018 Jul 1;25(7):800-808. doi: 10.1093/jamia/ocy013. PMID: 29584896; PMCID: PMC6658868.
20. Stevenson M, Guo Y. Disambiguation in the biomedical domain: the role of ambiguity type. J Biomed Inform. 2010 Dec;43(6):972-81. doi: 10.1016/j.jbi.2010.08.009. Epub 2010 Sep 9. PMID: 20816855.
21. McInnes BT, Stevenson M. Determining the difficulty of Word Sense Disambiguation. J Biomed Inform. 2014 Feb;47:83-90. doi: 10.1016/j.jbi.2013.09.009. Epub 2013 Sep 26. PMID: 24076369.
22. Eneko Agirre, Aitor Soroa, Mark Stevenson, Graph-based Word Sense Disambiguation of biomedical documents, Bioinformatics, Volume 26, Issue 22, November 2010, Pages 2889–2896,
23. Garla VN, Brandt C. Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification. J Am Med Inform Assoc. 2013 Sep-Oct;20(5):882-6. doi: 10.1136/amiajnl-2012-001350. Epub 2012 Oct 16. Erratum in: J Am Med Inform Assoc. 2014 May-Jun;21(3):568. PMID: 23077130; PMCID: PMC3756260.
24. Jimeno Yepes A. Word embeddings and recurrent neural networks based on Long-Short Term Memory nodes in supervised biomedical word sense disambiguation. J Biomed Inform. 2017 Sep;73:137-147. doi: 10.1016/j.jbi.2017.08.001. Epub 2017 Aug 7. PMID: 28797709.