International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 6 Volume No.6 October - November – 2022
DOI:10.46647/ijetms.2022.v06i06.086 ISSN: 2581-4621

A SURVEY ON BIOMEDICAL TEXT DOCUMENT CLASSIFICATION

Mr. D Krishna[1], Erukulla Laasya[2], A Sowmya Sri[3], T Ravinder Reddy[4], Akhil Sanjoy[5]
Associate Professor, Department of Computer Science and Engineering[1]
IV B.Tech Students, Department of Computer Science and Engineering[2,3,4,5]
ACE Engineering College, Hyderabad, Telangana, India

ABSTRACT
Information extraction, information retrieval, and text classification are only a few of the important study areas that fall under the heading of "bio medical text classification." In order to increase understanding of various information extraction opportunities in the field of data mining, this study analyses several text categorization approaches used in practise, their strengths and shortcomings. We have gathered a dataset with a strong emphasis on three categories, including "Thyroid Cancer," "Lung Cancer," and "Colon Cancer." This essay offers an empirical investigation of a classifier. Benchmarks for biomedical text were used to conduct the experiment. We study many metaheuristic algorithms, including genetic algorithms, particle swarm optimization, firefly, cuckoo, and bat algorithms. The suggested multiple classifier system also outperforms ensemble learning, ensemble pruning, and conventional classification algorithms.In the data we use predict the Biomedical text document classification is whether it's Thyroid Cancer, Lung Cancer, Colon Cancer based on the performed basic EDA, text pre-processing, build different models, such as LogisticRegression,DecisiontreeClassification,RandomForestClassification.
Keywords: *Biomedical Text, Classifier, Categorization, Metaheuristic Algorithms, Genetic Algorithm, Swarm Optimization.*

INTRODUCTION
An important source of data for biomedical research may be found in the vast amount of biomedical text. Biological text documents, such as scholarly publications, biomedical databases, and case reports, are characterised by a tremendous amount of unstructured and sparse information. Text mining uses tools and methods from several fields, including machine learning, information retrieval, and computational linguistics, to extract useful information from unstructured text. One of the most promising techniques in the biomedical field that has generated a lot of study attention is text mining. Text mining in the biomedical field has a wide range of successful applications, including the discovery of disease-specific knowledge, cancer diagnosis, treatment, and prevention, patient obesity status determination, risk factor identification for heart disease, annotation of gene expression, and the discovery of drug targets and candidates.
Biomedical text mining employs the same phases as other areas' text processing (specifically, format dialogue, tokenization, stop word removal, normalisation, stemming, dictionary creation, and vector space construction). One critical difficulty in developing reliable categorization methods for text documents is determining an acceptable representation model for the content. Due to its simple structure, the vector space model (also known as term vector model) is one of the most often used representation systems for processing text texts. Each text document is represented in this approach as a vector of IDs (index terms).High dimensional feature space, relevance, and feature sparsity are problems for the vector space model. Words are treated as statistically independent since each document is represented as a bag of words with the appropriate frequencies. Consequently, word order is not taken into account.
As new discoveries are discovered and procedures evolve, clinical knowledge is continually developing. This information is vital, but it is frequently buried in text in a variety of sources, such as journal papers and clinical narratives in the patient record. Through automated methodologies,

DOI:10.46647/ijetms.2022.v06i06.086 ISSN: 2581-4621

disease-related information may be retrieved and combined from these different text sources to better understand disease features (e.g., treatment or symptoms) and how they may evolve over time.
Text categorization is the process of assigning one or more categories to natural language documents from a present collection. Predefined categories are often thematic, however there are applications where categories are established by other criteria, such as genre categorization, email classification by priority, and so on. Text categorization is an example of supervised learning in which a set of categories and samples of texts from those categories are provided. This study will not address challenges of unsupervised learning, often known as text clustering, in which the categories are unknown in advance.
Biomedical electronic document databases are rapidly developing, culminating in massive digital archives. Manually organising and searching these papers is becoming increasingly expensive and time demanding. Because of its fast expansion, biomedical literature has been the focus of extensive information retrieval and machine learning research during the last few decades. Text categorization is a difficult study topic in which text documents are classified using predetermined labels depending on their content. Improving text categorization approaches for biomedical databases is critical for overcoming information overload and making indexing, filtering, and maintaining the expanding number of articles in those databases easier.
The majority of text classification methods now in use describe texts as vectors. Key elements and concepts are extracted from text and used as features in the vector space model. The absence of semantic links between important items and concepts in the text is a drawback of the vector format. Recently, methods for modelling complicated data, such as protein sequences and structures, social networks, and graphs, have started to gain favour. The use of "rich" semantic representation of relationships between important entities and concepts in a text is a benefit of graph modelling, which may lead to better classification outcomes.
There is an increasing amount of biological material available on the internet these days. The requirement to filter and categorise it in order to make the most use of the data is becoming a significant difficulty. There are several studies on text categorization (or classification) based on word occurrences, most of which come from the machine learning discipline. On the contrary, in this study, we suggest that for classification, we integrate the linguistic and structural aspects of texts with the absence/presence of patterns in documents. The findings of a prior study in which we categorised biomedical publications using integrated multi-scale descriptors prompted us to pursue this notion.

LITERATURE REVIEW

In the data we use predict the Biomedical text document classification is whether it's Thyroid Cancer, Lung Cancer, Colon Cancer based on the performed basicEDA, text pre-processing, build different models, such as Logistic Regression, Decisiontree Classification, RandomForest Classification, XGBboost Classifier.
[1] Divina, Federico, Onan, Aytug – This article outlines a brand-new biomedical text classification method based on diversity-based ensemble pruning and swarm-optimized latent Dirichlet allocation. Due to the vast amount of unstructured information accessible, biomedical text classification is a crucial study area. A common representation method for text texts is the latent Dirichlet allocation (LDA), which performs better than other linguistic representation methods like latent semantic analysis and probabilistic latent semantic analysis. They discovered that the effectiveness of LDA greatly depends on the determination of acceptable parameter values. Additionally, it has been empirically confirmed that using metaheuristic optimization techniques to calibrate the LDA's parameters produces positive outcomes for categorising biomedical literature.In addition, an assembly trimming strategy based on integrated diversity measurements is used in the current text classification scheme to discover a robust multiple classifier system with good predictive performance.

DOI:10.46647/ijetms.2022.v06i06.086 ISSN: 2581-4621

[2] E. S. Chen, G. Hripcsak, H. Xu, M. Markatou, and C. Friedman – The work in this paper is focused on a strategy for gathering disease-specific knowledge, as well as a feasibility study The strategy is based on using a mix of natural language processing (NLP) and statistical techniques to biological and clinical texts. Based on the patient record, the technique permitted the extraction of knowledge about the pharmaceuticals practitioners use for patients with certain disorders, as well as knowledge of drugs regularly involved in controlled trials for those same diseases. They found the findings to be suitable when comparing the disease-drug linkages: the two text sources contained consistent as well as complementary knowledge, and a manual assessment of the top five disease-drug associations by a medical expert validated their accuracy across the disorders.

[3] R. Rodriguez-Esteban -This lesson is meant for biologists and computational biologists who want to expand their bioinformatics arsenal using text mining methods. The lesson addresses the link between progressive multifocal leukoencephalopathy (PML) and antibodies as an example. Recent instances of PML have been linked to the use of monoclonal antibodies such as efalizumab. Those looking for a more in-depth introduction to text mining may wish to read other evaluations. Understanding vast quantities of text with a computer is more difficult than merely providing a computer with a grammar and a dictionary. To interpret language, a machine, like a person, need specific expertise.

[4] Meenakshi Mishra, Jun Huan, Said Bleik, Min Song – The research of this paper is they look at a text graph representation for the problem of text classification. They discover high level ideas derived from a database of restricted biomedical phrases and develop a rich graph structure that encompasses significant concepts and interactions in our representation. This technique guarantees that graphs are described using a consistent terminology, allowing for easier comparison. The document graphs are then classified using a set-based graph kernel that is intuitively understandable and capable of dealing with the disconnectedness of the created idea graphs. Thiscompare our method to others that use non-graph, text-based characteristics. This also run a comparison of different kernels to discover which performs better.

[5] Zerida, Nadia Lucas, Nadine Crémilleux, Bruno -In this research, they presented a new technique of categorising biomedical publications based on two novel assumptions. On the one hand, they mixed linguistic, structural, and metric characteristics to create data mining patterns. On the other hand, they also used an exclusion-inclusion strategy to account for the relevance of the lack of patterns to the classification job. To prevent a sharp difference between the absence and presence of a pattern, the exclusion-inclusion technique employs two regret measures to quantify the interest of a weak pattern in comparison to other classes and among patterns within the same class. The global decision is based on the generalisation of the local patterns, first by utilising patterns that exclude classes, and secondly on the regret ratios.

[6] Minsuk Lee, Weiqing Wang and Hong Yu -This work provides the first and most advanced topic identification algorithms in biomedical writings. The assessment concluded that supervised topic spotting provides the best performance for topic detection in OMIM data. Our results reveal that, while unsupervised topic clustering approaches perform worse than topic spotting methods, they perform much better than the baseline. Furthermore, when used to detecting subjects described by biologists in their review papers, the performance of topic clustering algorithms improves. Our findings demonstrate that topic clustering algorithms are capable of dealing with real-world occurrences. The findings also show that the performance of topic clustering improves with data quality.

[7]Kiritchenko, Svetlana – The study finds that they have investigated two major features of hierarchical categorization in this dissertation: learning algorithms and performance evaluation. They presented the concept of consistent hierarchical categorization, which helps end-users understand and comprehend classification findings. Only a local top-down technique achieves consistent categorization among the previously presented hierarchical learning algorithms. This approach is extended in this paper to the general situation of DAG class hierarchies and possibly internal class assignments. A novel global hierarchical strategy for conducting consistent categorization is also

DOI:10.46647/ijetms.2022.v06i06.086 ISSN: 2581-4621

proposed.This is a broad framework for transforming a traditional "flat" learning algorithm into a hierarchical learning algorithm. An comprehensive series of tests on actual and synthetic data show that the suggested strategy outperforms both the related "flat" and the local top-down methods. They have utilised a unique hierarchical evaluation measure that, according to a variety of formal criteria, outperforms existing hierarchical and non-hierarchical assessment procedures.

[8] CyrilleYetuYetuKesiku , Andrea Chaves-Villota and Begonya Garcia-Zapirain -This paper explores the different difficulties in the process of classifying biomedical texts by concentrating on a number of issues, identifying the structure of biomedical data forclassification task, difficulty of method performance, enumerating the numerous issues and difficulties that the text classification which might address in the biomedical domain and at last to asses biomedical test classification test the most often employed metrics. Reviewing the numerous publications selected for this study's evaluation led us to two important problems related to the methods used for biomedical text categorization. The most common transfer learning method that uses pre-trained algorithms uses a dataset of wide text classification settings, which explains the data-centric approach.

[9] Manirupa Das, Juanxi Li, Eric Fosler-Lussier, Simon Lin, Steve Rust, Yungui Huang and Rajiv Ramnath – The study finds that by training document representations without the use of any external supervision labels for unsupervised semantic tagging of a huge collection of documents, they created a unique sequence-to-set end-to-end encoder-decoder-based neural framework for multi-label prediction. They also discovered that a multi-term prediction training objective that jointly optimises both prediction of the TFIDF-based document pseudo-labels and the log likelihood of the labels given the document encoding, outperforms conventional approaches in this unsupervised task setting of PRF-based semantic tagging for query expansion.

[10] Man LAN, Chew Lim TAN, Jian SU, Hwee Boon LOW -This study looked at various term weighting and term type representations for categorising bio-medical texts. The performance of classification is enhanced by term weighting techniques. In particular, this study and the newswire domain demonstrate the classification power of our suggested tf.rf approach. However, compared to the bag-of-words method, named entity-based representations are not much better. This bolsters the overall finding that straightforward NLP-based representations do not enhance text classification performance. NLP approaches need to make a lot of progress, according to the researchers, before they can be used to enhancetext categorization

Authors	Title	Year of Publication	Methods
Divina, Federico, Onan, Aytug	Biomedical Text Categorization Based on Ensemble Pruning and Optimized Topic Modelling	2018	The Latent Semantic Analysis Method Latent Dirichlet Allocation Ensemble Generation Ensemble Pruning
Meenakshi Mishra, Jun Huan, Said Bleik, Min Song	Biomedical Text Categorization with Concept GraphRepresentations Using a Controlled Vocabulary	2012	Graph Construction and Processing Classifier learning with Kernals
Man LAN, Chew Lim TAN, Jian SU, Hwee Boon LOW	Text Representations for Text Categorization: A Case Study inBiomedical Domain	2007	Named Entity Term Weighting Methods Traditional Method

DOI:10.46647/ijetms.2022.v06i06.086 ISSN: 2581-4621

E. S. Chen, G. Hripcsak, H. Xu, M. Markatou, and C. Friedman	Automated acquisition of disease–drug knowledge from biomedical and clinical documents: an initial study	2008	Randomized Controlled Trails BioMedLEE and MedLEE for extracting UMLS
Minsuk Lee, Weiqing Wangand Hong Yu	Exploring supervised and unsupervised methods to detect topics inbiomedical text	2006	OMIM Statistics Topic Spotting Topic Clustering Naïve Bayes Single-Pass Clustering
Manirupa Das, Juanxi Li, Eric Fosler-Lussier, Simon Lin, Steve Rust, Yungui Huang and Rajiv Ramnath	Sequence-to-Set Semantic Tagging for Complex Query Reformulation andAutomated Text Categorization in Biomedical IR using Self-Attention	2020	Sequence-to-Sequence Modelling Approaches for Text Classification
Zerida, Nadia Lucas, Nadine Crémilleux, Bruno	Exclusion-inclusion based text categorization of biomedical articles	2007	Exclusion-Inclusion Method (Hypothesis)

CONCLUSION

The reason for this research is to classify the biomedical text using some heuristic algorithms. It is developed using Machine Learning. Text mining is a significant study area that encompasses various fields, including information retrieval, information extraction, and text classification.Based on the data, we forecast if it is Thyroid Cancer, Lung Cancer, or Colon Cancer using basic EDA, text pre-processing, and several models such as Logistic Regression, Decision Tree Classification and Random Forest Classification.

ACKNOWLEDGEMENT

We owe a sincere thanks to Mr.D.Krishna and Mrs.Soppari Kavitha who served as our guides for theirkind cooperation and guidance,especially our head of the department Dr.M.VijayaSaradhi , is greatlyappreciated for his time and guidance which helped us in the completion of this project which wouldhave seemed difficult without their constant support and valuable suggestions.

REFERENCES

[1] Divina, Federico, Onan, Aytug –"Biomedical Text Categorization Based on Ensemble Pruning and Optimized Topic Modelling", 2018
[2] E. S. Chen, G. Hripcsak, H. Xu, M. Markatou, and C. Friedman, "Automated acquisition of disease–drug knowledge from biomedical and clinical documents: an initial study," Journal of the American Medical Informatics Association, vol. 15, no. 1, pp. 87–98, 2008.
[3] R. Rodriguez-Esteban, "Biomedical text mining and its applications," PLoS Computational Biology, vol. 5, no. 12, Article ID e1000597, 2009.
[4] Meenakshi Mishra, Jun Huan, Said Bleik, Min Song – "Biomedical Text Categorization with Concept GraphRepresentations Using a Controlled Vocabulary", 2012.

DOI:10.46647/ijetms.2022.v06i06.086 ISSN: 2581-4621

[5] Zerida, Nadia Lucas, Nadine Crémilleux, Bruno– "Exclusion-inclusion based text categorization of biomedical articles", 2007
[6] Minsuk Lee, Weiqing Wangand Hong Yu – "Exploring supervised and unsupervised methods to detect topics inbiomedical text", 2006
[7] Kiritchenko, Svetlana –"Hierarchical text categorization and its application to bioinformatics", 2005
[8] CyrilleYetuYetuKesiku, Andrea Chaves-Villota and Begonya Garcia-Zapirain – "Natural Language Processing Techniques for Text Classification of Biomedical Documents: A Systematic Review", 2022
[9] Manirupa Das, Juanxi Li, Eric Fosler-Lussier, Simon Lin, Steve Rust, Yungui Huang and Rajiv Ramnath – "Sequence-to-Set Semantic Tagging for Complex Query Reformulation andAutomated Text Categorization in Biomedical IR using Self-Attention", 2020
[10] Man LAN, Chew Lim TAN, Jian SU, Hwee Boon LOW – "Text Representations for Text Categorization: A Case Study inBiomedical Domain", 2007