International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 3 Volume No.7 May - June – 2023
DOI:10.46647/ijetms.2023.v07i03.121 ISSN: 2581-4621

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, retrieval, and text categorization are only a few of the significant research fields covered by "bio medical text classification." This study examines many text categorization techniques utilised in practise, as well as their strengths and weaknesses, in order to improve knowledge of various information extraction opportunities in the field of data mining. We compiled a dataset with a focus on three categories: "Thyroid Cancer," "Lung Cancer," and "Colon Cancer." This paper presents an empirical study of a classifier. The investigation was carried out using biomedical literature benchmarks. Many metaheuristic algorithms are investigated, including genetic algorithms, particle swarm optimisation, firefly, cuckoo, and bat algorithms. In addition, the proposed multiple classifier system outperforms ensemble learning, ensemble pruning, and traditional classification methods. Based on the data, we forecast if it is Thyroid Cancer, Lung Cancer, or Colon Cancer using basic EDA, text preprocessing, and several models such as Logistic Regression, Decision Tree Classification, and Random Forest Classification.

INTRODUCTION
The vast amount of biomedical text is an important source of data for biomedical research. There is a lot of unstructured and sparse information in biological text documents like case reports, scholarly publications, and biomedical databases. To extract useful information from unstructured text, text mining employs tools and techniques from machine learning, information retrieval, and computational linguistics. Text mining is one of the biomedical field's most promising methods that has received a lot of research attention. The discovery of disease-specific knowledge, cancer diagnosis, treatment, and prevention, determination of patient obesity status, risk factor identification for heart disease, annotation of gene expression, and the discovery of drug targets and candidates are just a few of the many successful applications of text mining in the biomedical field.
Biomedical text mining utilizes similar stages as other regions' text handling (in particular, design exchange, tokenization, stop word evacuation, standardization, stemming, word reference creation, and vector space development). One basic trouble in creating dependable classification strategies for message reports is deciding an adequate portrayal model for the substance. Because of its basic design, the vector space model (otherwise called term vector model) is one of the most frequently involved portrayal frameworks for handling text texts. Every text report is addressed in this methodology as a vector of IDs (list terms). High layered include space, significance, and component sparsity are issues for the vector space model. Words are treated as measurably free since each report is addressed as a pack of words with the suitable frequencies. Thus, word request isn't considered.
The process of classifying natural language documents from a current collection into one or more categories is known as text categorization. While predefined categories are typically thematic, there are some applications where categories are established based on other criteria, such as email classification based on priority or genre. One example of supervised learning is text categorization, in which a set of categories and examples of texts from those categories are provided. The difficulties

DOI:10.46647/ijetms.2023.v07i03.121 ISSN: 2581-4621

of unsupervised learning, also known as text clustering, in which the categories are not known in advance, will not be addressed in this study.

EXISTING SYSTEM

From rule-based to machine-learning-based systems, numerous biomedical text document classification systems are currently in use. One illustration is:

MeSH : Medical Subject Headings is a popular indexing and classification system for biomedical literature. The National Library of Medicine (NLM) developed the Medical Subject Headings (MeSH) Classifier system. To assign MeSH terms to biomedical articles, it makes use of machine learning techniques like Naive Bayes and Support Vector Machines. To perform the classification, the classifier makes use of features like word frequencies, MeSH term cooccurrence, and citation information.

As a result, the current system has a few drawbacks:

- There is less data available
- There isn't enough accuracy
- There isn't enough quality control

PROPOSED SYSTEM

Classification algorithms are utilized in the proposed system for the classification of biomedical text documents. The following steps need to be taken:

- Data preparation
- Feature extraction
- Training and testing
- Metric accuracy-based performance evaluation
- System optimization

Advantages :

- **Effective Information Acquisition:** Classification of biomedical text documents aids in the organization of large volumes of biomedical literature and makes it simpler to quickly retrieve specific information. Researchers and medical professionals can efficiently search for relevant articles, research papers, or clinical guidelines in their field of interest by categorizing documents into various classes or topics.
- **Developing New Knowledge:** By highlighting connections and patterns among various biomedical concepts, biomedical text document classification facilitates knowledge discovery. Researchers can gain insight into complex biomedical phenomena and uncover hidden connections by categorizing documents into specific topics or domains.

This may result in new discoveries, enhanced knowledge of diseases, and enhanced patient care.

- **Recommendations for Personalized Medicine and Treatment:** By assisting in the identification of suitable treatment options based on patient characteristics and medical conditions, the classification of biomedical text documents can contribute to personalized medicine. Healthcare providers can make more informed decisions regarding treatment plans, drug prescriptions, and therapeutic interventions by analyzing a patient's medical records and classifying them into specific categories or disease subtypes.

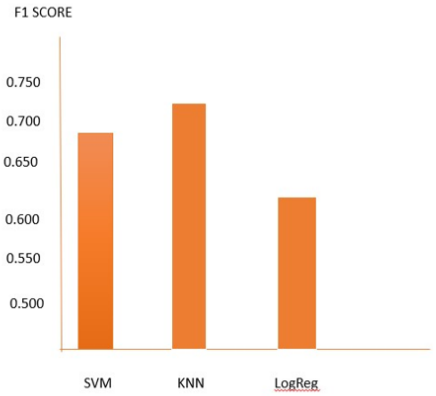
PERFORMANCE METRICS

A number of important metrics are frequently used to assess a biomedical text classification model's performance. The model's ability to correctly classify documents is revealed by these metrics. Among the most frequently used evaluation metrics are:

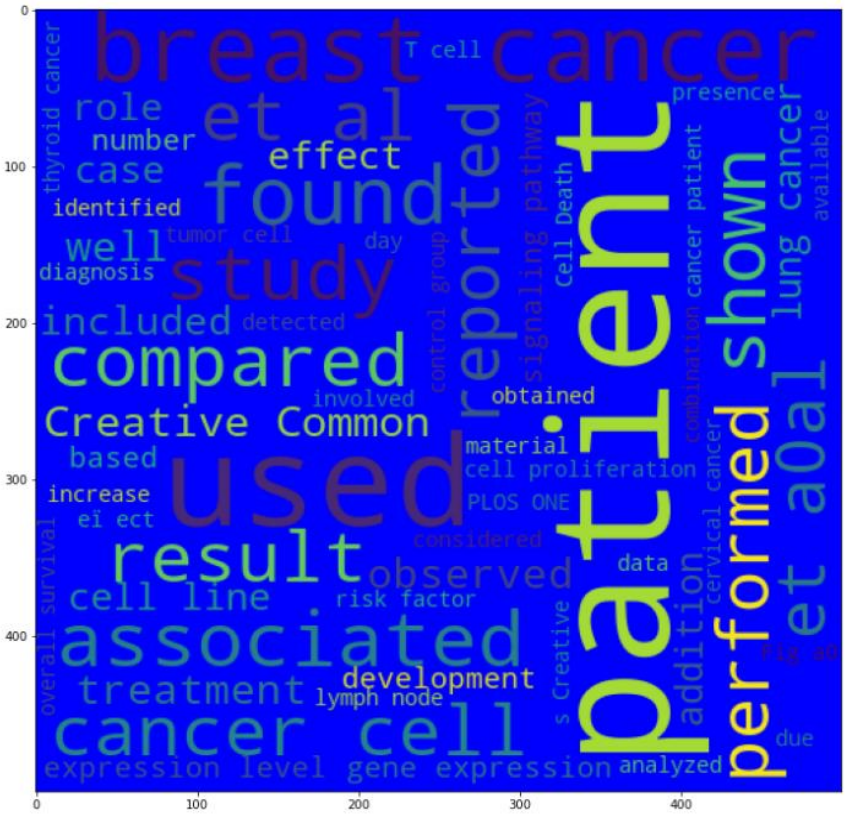
Accuracy: It shows the proportion of correctly classified documents out of the total number of documents, indicating the overall correctness of the model's predictions.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 3 Volume No.7 May - June – 2023
DOI:10.46647/ijetms.2023.v07i03.121 ISSN: 2581-4621

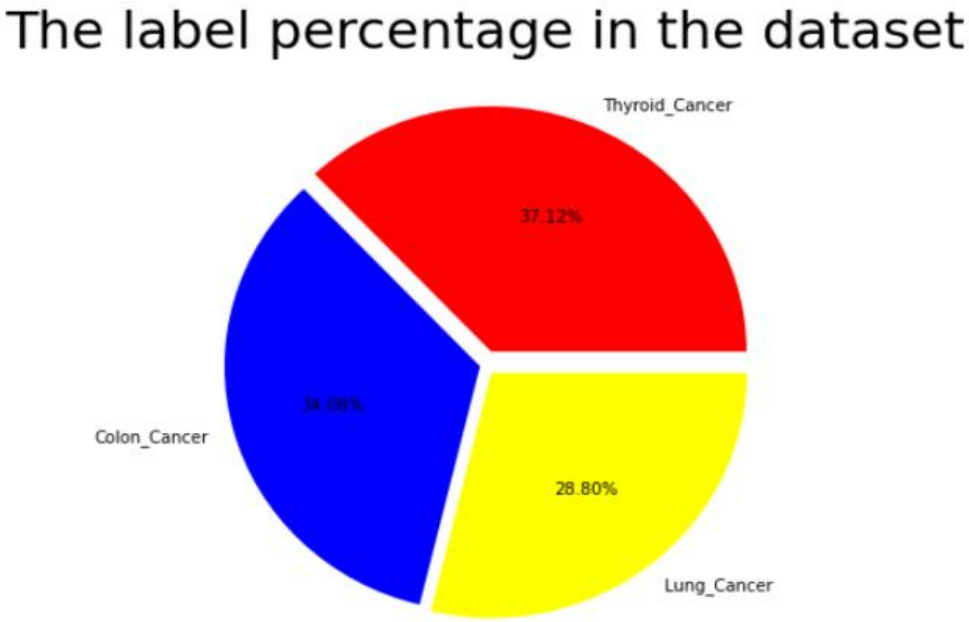
Precision: The model's ability to correctly identify positive instances within the predicted positive class is measured by precision. The ratio of true positive instances to the total of true positive and false positive instances is quantified by this method.
Recall: The ability of the model to identify all positive instances in the dataset is calculated using recall, which is also known as sensitivity. It is the ratio of instances that are true positives to the total of instances that are true positives and false negatives.
F1 Score: The F1 score provides a balanced evaluation of the model's performance by combining precision and recall into a single metric. It is the harmonic mean of precision and recall, which is calculated as 2 * (precision minus recall) / (precision minus recall). When there is a disparity between the classes in the dataset, the F1 score is frequently utilized.
In biomedical text classification, additional specialized evaluation techniques are frequently utilized because of the importance of particular classes and the presence of imbalanced datasets. These methods include confusion matrices, precision-recall curves, class-specific metrics like specificity, and macro-averaged metrics.

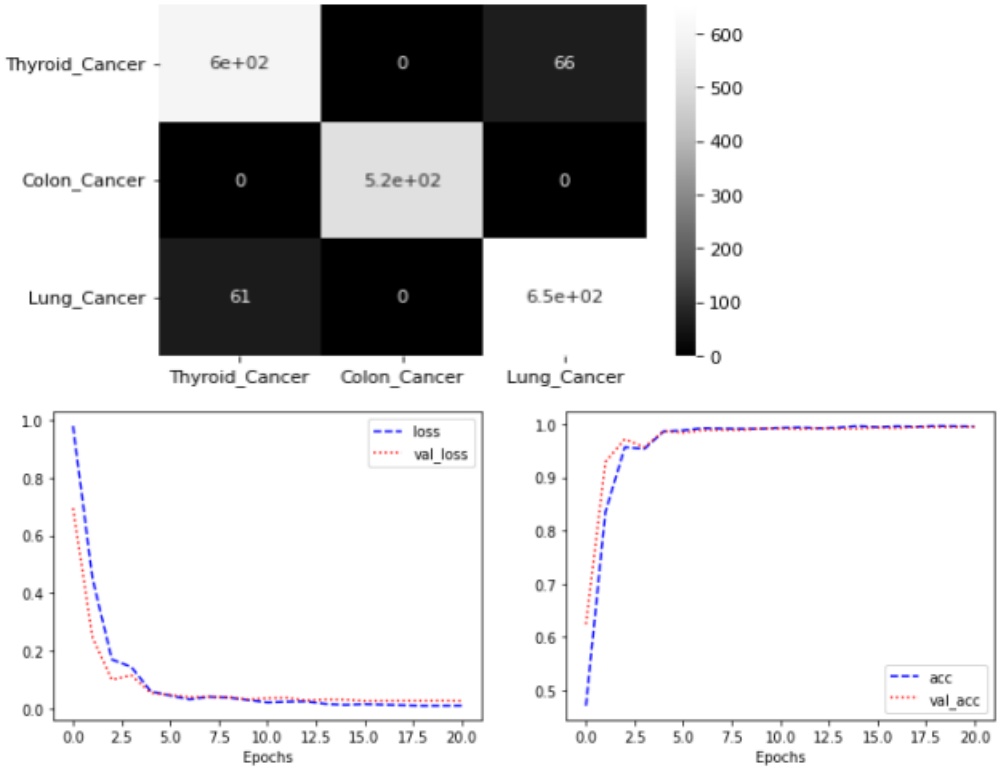


RESULTS



International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 3 Volume No.7 May - June – 2023
DOI:10.46647/ijetms.2023.v07i03.121 ISSN: 2581-4621





CONCLUSION

In conclusion, biomedical text document classification is an essential component of biomedical research because it enables researchers to quickly locate relevant documents for their investigations. Pre-processing, feature extraction, feature selection, classification, and assessment are all parts that

DOI:10.46647/ijetms.2023.v07i03.121 ISSN: 2581-4621

work together to classify a lot of text data in an accurate way. The application of deep learning, the creation of interpretable models, the integration of multiple modalities, and the improvement of data quality are just a few of the remaining obstacles. These issues point to possible future developments that have the potential to significantly improve the efficiency and accuracy of biomedical text document classification. With additional research and development, the topic of biomedical text document categorization has the potential to make it possible to process and analyze huge volumes of biomedical data with greater efficiency. This could lead to new insights and breakthroughs in biomedical research.

ACKNOWLEDGEMENTS

The survey's research would not have been successful without the direction of our guide, **Mr. D. Krishna**, and the project coordinator, **Dr. K. Prem Kumar**. Our sincere gratitude goes out to **Dr. M. V. Vijaya Saradhi**, Head of Computer Science and Engineering Department at ACE Engineering College, for his invaluable and steadfast assistance during the completion of this project.

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 Graph Representations Using a Controlled Vocabulary", 2012.

[5] Zerida, Nadia Lucas, Nadine Crémilleux, Bruno – "Exclusion-inclusion based text categorization of biomedical articles", 2007

[6] Minsuk Lee, Weiqing Wang and Hong Yu – "Exploring supervised and unsupervised methods to detect topics in biomedical text", 2006

[7] Kiritchenko, Svetlana – "Hierarchical text categorization and its application to bioinformatics", 2005

[8] Cyrille YetuYetu Kesiku, 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 and Automated 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 in Biomedical Domain", 2007