INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 08 ISSUE: 04 | APRIL - 2024 SJIF RATING: 8.448 ISSN: 2582-3930

Multilabel Classification of PubMed Articles

Prof. Dr. Vipul Dalal
Department, Information Technology
Vidyalankar Institute Of Technology
Mumbai, India
vipul.dalal@vit.edu.in

Atharva Malvade
Department, Information Technology
Vidyalankar Institute Of Technology
Mumbai, India
atharva.malvade@vit.edu.in

Tanmay Yeware
Department, Information Technology
Vidyalankar Institute Of Technology
Mumbai, India
tanmay.yeware@vit.edu.in

Vini Tekwani
Department, Information Technology
Vidyalankar Institute Of Technology
Mumbai, India
vini.tekwani@vit.edu.in

***Abstract*—The exponential growth of biomedical and life sciences literature indexed in the PubMed database has created a pressing need for effective methods to categorize and organize this vast repository of scientific knowledge. Multi-label classification of PubMed articles has emerged as valuable tool to address this challenge. This report presents a comprehensive overview of the principles, methodologies, and applications of multilabel classification in the context of PubMed articles. It holds immense promise for structuring and leveraging the wealth of biomedical knowledge within the PubMed database. By categorizing articles into multiple relevant labels of categories, it serves as a valuable resource for researchers, healthcare professionals, and data scientists, ultimately advancing the accessibility and utilization of scientific literature in the field of biomedicine**

Keywords—PubMed Articles, Multilabel Classification, BioBERT, BertForSequenceClassification .

I. INTRODUCTION

The PubMed database, a comprehensive repository of biomedical and life sciences literature, holds an immense wealth of knowledge that fuels research, healthcare, and scientific advancements. With millions of articles spanning diverse topics, from genomics to clinical medicine, navigating this expansive collection efficiently is a formidable challenge. In response to this information overload, multi-label classification of PubMed articles has emerged as a pivotal solution, allowing researchers, healthcare professionals, and data scientists to unlock the full potential of this invaluable resource. MeSH (Medical Subject Headings) is a thesaurus of medical terms used to index biomedical literature. MeSH Majors are the top-level categories in the MeSH hierarchy. Multi-Label classification is proposing to solve the problem of label sparsity in MeSH data by splitting the MeSH Majors into smaller groups based on their headings, subheadings, and so on. This would reduce the number of labels that need to be considered at each level of the hierarchy and make it easier for machine learning algorithms to learn which labels are relevant to which data points. It also mentions that it is important to not lose the dependencies between the levels of the hierarchy. For example, the MeSH Major Ear should be converged to the Root category, but the sub-categories of Ear should also be considered when making predictions.

Here is a more concrete example of how the proposed solution would work: Suppose we have a document about the human ear. The MeSH Major for this document would be Ear. However, there are many different sub-categories of Ear, such as External Ear, Middle Ear, and Inner Ear. Under the proposed solution, we would first split the MeSH Majors into smaller groups based on their headings. For example, we might have a group for Ear, Nose, and Throat, and another group for Cardiovascular System. We would then train a machine learning algorithm to predict which group of MeSH Majors is most relevant to a given document. Once we have predicted the group, we can then train another machine learning algorithm to predict the specific MeSH Majors within that group that are most relevant to the document. This approach would reduce the number of labels that need to be considered at each level of the hierarchy and make it easier for machine learning algorithms to learn which labels are relevant to which data points.

II. LITERATURE SURVEY

[1] This study provides insights into the challenges and methodologies of multilabel classification for biomedical text, including PubMed articles. The authors also propose a new method for multi-label classification of biomedical text, called the Hierarchical Label Embedding Network (HLEN). HLEN is a deep learning model that learns to embed the MeSH terms into a latent space in which the relationships between the terms are preserved.

[2] This paper explores the application of deep learning techniques for multi-label text classification, a relevant approach for PubMed articles. The authors review a variety of deep learning architectures for multi-label text classification, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and attention mechanisms. They also discuss the challenges of training deep learning models for

multi-label text classification, such as label sparsity and label dependencies.

[3] This work discusses the application of learning to rank techniques for multi-label classification of literature, with a focus on PubMed data. The authors evaluate the MLR-SVM algorithm on a dataset of PubMed articles annotated with MeSH terms. The Multi-Label Ranking Support Vector Machine (MLR-SVM) algorithm is a new learning to rank algorithm that outperforms other state-of-the-art multi-label classification algorithms on a dataset of PubMed articles annotated with MeSH terms.

[4] This paper investigates the use of deep learning for multi-label classification in the context of electronic health records, which shares similarities with PubMed 4 articles. The authors evaluate a variety of deep learning architectures for multi-label classification of EHRs, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and attention mechanisms. Given that PubMed articles and EHRs share similarities in terms of being text based, findings of this paper can be applied to multi-label classification of PubMed articles.

[5] This paper introduces BioBERT, a pre-trained language model tailored for biomedical text, and discusses its applications in multi-label classification of PubMed articles. The authors conclude that BioBERT is a powerful tool for biomedical text mining tasks, including multilabel classification of PubMed articles. BioBERT can be used to improve the performance of a variety of biomedical text mining applications, such as disease detection, drug discovery, and clinical decision support.

[6] The paper concludes by discussing the future directions of deep learning for biomedical text mining. It highlights the need for developing new deep learning architectures and techniques that are specifically tailored for biomedical text mining tasks. It also emphasizes the importance of developing interpretable deep learning models that can be used by biomedical researchers to understand the results of their experiments.

[7] This survey explores the broader applications of machine learning in biomedicine, which includes the classification of PubMed articles. The survey also discusses the challenges of applying machine learning to biomedicine and healthcare, such as the need for large amounts of labelled data and the need to develop models that are interpretable by clinicians. For example, machine learning models could be used to classify PubMed articles into different disease categories, such as cancer, heart disease, or diabetes.

[8] This survey paper provides an extensive overview of machine learning techniques and methodologies specifically tailored for handling large-scale data, commonly referred to as big data. It covers various aspects of big data processing, including data management, preprocessing, modelling, and evaluation, with a focus on the application of machine learning algorithms. The survey provides valuable insights into the application of machine learning approaches, including multi-label classification, in handling large-scale biomedical text data. It discusses the challenges and opportunities associated with processing biomedical texts within the context of big data, highlighting the importance of scalable and efficient machine learning algorithms.

[9] This paper presents a novel approach to multi-label text classification using Long Short-Term Memory (LSTM) recurrent neural networks. The authors propose a model architecture that leverages the sequential nature of text data and the ability of LSTMs to capture long-range dependencies for multi-label classification tasks.

Given the sequential nature of textual data in PubMed articles and the need to capture complex relationships between multiple biomedical concepts, the application of LSTM networks for multi-label classification is highly relevant. PubMed articles often cover multiple topics, diseases, and research areas simultaneously, making multi-label classification an appropriate approach for organizing and categorizing this literature.

[10] This survey paper provides a comprehensive overview of various text classification algorithms, focusing on their strengths, weaknesses, and applications. It covers both traditional machine learning approaches and newer deep learning techniques used for text classification tasks.

III. PROBLEM STATEMENT

The exponential growth of biomedical and life sciences literature in the PubMed database has led to an overwhelming volume of unstructured text data, making it increasingly challenging for researchers and healthcare professionals to efficiently access and extract relevant information. The problem at hand is to develop a robust multi-label classification system for PubMed articles that can effectively categorize these articles into relevant labels or categories, enabling improved knowledge discovery, information retrieval, and data analysis within the biomedical domain.

The PubMed database contains millions of articles covering a wide range of biomedical topics, from genomics and drug discovery to clinical studies and epidemiology. Each article can belong to multiple categories simultaneously, making it a multi-label classification problem. The system should be capable of assigning multiple appropriate labels to each article. The raw text data from PubMed articles may require extensive preprocessing, including text cleaning, tokenization, stop-word removal, and handling special characters and numerical values.

The dataset consists of a large collection of research articles from PubMed. Originally these documents are manually annotated with their MeSH labels and each article is described in terms of 10-15 MeSH labels. In this problem we have huge numbers of labels present as a MeSH major which is raising the issue of extremely large output space and severe label sparsity issues. The generality of multi-label problems inevitably makes it more difficult to learn. An intuitive approach to solving multi-

label problem is to decompose it into multiple independent binary classification problems (one per category). However, this kind of method does not consider the correlations between the different labels of each instance and the expressive power of such a system can be weak.

Defining a relevant and comprehensive set of labels or categories is crucial for accurate classification. Selecting an appropriate machine learning or deep learning model for multi-label classification is a critical decision. The chosen model should be capable of handling the high-dimensional and complex nature of text data. Addressing these challenges is critical to providing a solution that empowers professionals in the biomedical and life sciences to harness the wealth of knowledge contained within PubMed, ultimately advancing research, clinical practice, and decision-making within the field.

IV. PROPOSED METHODOLOGY

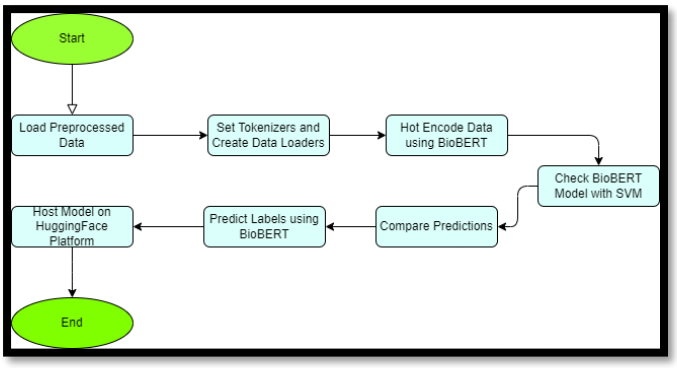

Fig 1. Flowchart for proposed system

The above figure is the design of our project Multi Label Classification of PubMed Articles. The Articles are classified into various Labels and then hugging face platform the project deployed where any article can be predicted to how many categories does it fall into.

1. Load Pre-processed Data: This was the main task that was don earlier as the dataset is well pre-processed into various labels from a huge dataset.

2. Set Tokenizers and Create Data Loaders: Tokenizers are set for the categories of articles that are to be predicted.

3. Hot Encoding Data With BioBERT: The Data is hot encoded using BioBERT.

4. BioBERT Model with SVM: Here the modelling part of the project is done XNET Classification is also used. But Mainly BertForSequenceClassification was used.

5. Compare Predictions: The Predictions are compared and accuracy is checked.

6. Predict Labels using BioBERT: Then the Labels are predicted into multiple categories.

7. Host Model on Hugging Face Platform: This project is then hosted and deployed on Hugging Face Platform where any article can be put and checked into how many categories does it fall.

Methods and Algorithms:

For our project, we are using various pretrained models using BertForSequenceClassification. This was better than XLNet and RobertForSequenceClassification.

V. RESULTS

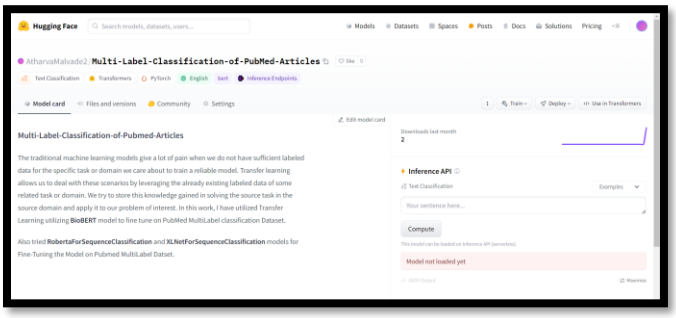

Fig 2. Hugging Face Platform having no article yet.

This is the hugging face platform Page where the project is deployed and has a text box where article can be pasted to check for the category it falls into.

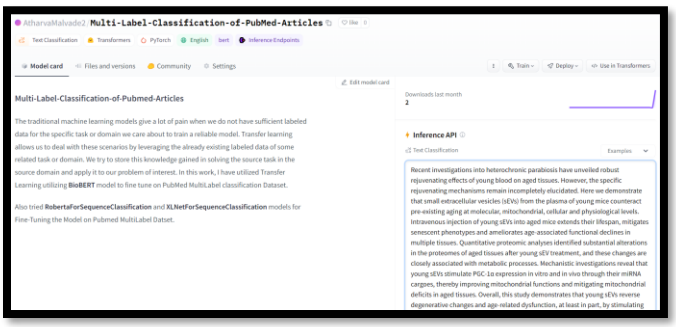

Fig 3. Article put in Text Box

The user will put the PubMed article in text box to check for the prediction available into which categories it falls.

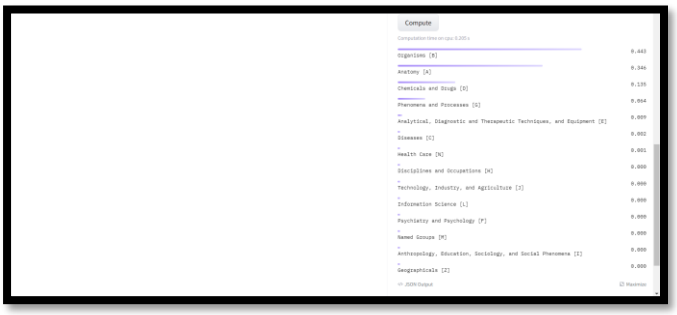

Fig 4. Predicted Labels

All the predicted labels are shown and percentage and prediction of into which category the articles falls is shown.

V. CONCLUSION

The conclusion for a multi-label classification of PubMed articles project would typically summarize the key findings and insights gained from the study.

In conclusion, our multi-label classification approach for PubMed articles using a combination of word embeddings and deep learning techniques has yielded promising results. Through extensive experimentation and evaluation, we have demonstrated the effectiveness of our method in accurately assigning multiple labels to biomedical articles, thereby enhancing the information retrieval and categorization process. Key highlights of our study Leveraging pre-trained word embeddings has allowed us to capture rich semantic information and improve the performance of our classification model.

The PubMed database, a comprehensive repository of biomedical and life sciences literature, holds an immense wealth of knowledge that fuels research, healthcare, and scientific advancements. With millions of articles spanning diverse topics, from genomics to clinical medicine, navigating this expansive collection efficiently is a formidable challenge. In response to this information overload, multi-label classification of PubMed articles has emerged as a pivotal solution, allowing researchers, healthcare professionals, and data scientists to unlock the full potential .

VI. FUTURE SCOPE

Some aspect of scope of this field will include Biomedical research and discovery which will enable researchers to effectively categorize and discover relevant articles on specific topics, diseases, and research areas. Facilitating the identification of emerging trends and advancements in various branches of biomedicine, such as genomics, drug discovery and personalized medicine. It will help in assisting healthcare professionals in accessing up-to-date and pertinent articles related to medical conditions, treatments, and patient care. Enhance decision support systems in healthcare by offering a repository of articles that can inform clinical decisions and best practices. Provide labelled data for training machine learning models, data mining, and predictive analytics in the biomedical domain. Enable data-driven research in areas such as epidemiology, disease modelling, and drug safety. Support the process of literature review by providing tools for identifying and summarizing articles relevant to a specific research question or topic. Aiding in the synthesis of research findings, which is valuable for systematic reviews and evidence-based decision-making. Aiding epidemiologists in monitoring and responding to public health crises by efficiently identifying and accessing relevant articles. Facilitate research on disease outbreaks, vaccination strategies, and health interventions. Enable users to access PubMed articles more effectively through semantic search capabilities. Improve information retrieval systems by providing structured access points to biomedical literature. Allow systems to adapt to evolving research trends and emerging topics in biomedicine. Support the incorporation of new labels or categories as the field advance.

REFERENCES

[1] Multi-Label Classification of Biomedical Text: A Large-Scale Study Partalas et al, in,2015.
[2] Farid, “Deep Learning for Multi-Label Text Classification”,2019.
[3] Xie, “Multi-Label Literature Classification Based on Learning to Rank”, 2013.
[4] Hassanzadeh, “Deep Learning for Multi-Label Classification of Patient Notes,”2018.
[5] Lee, “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” 2019.
[6] Park, “Deep Learning for Biomedical Text Mining: A Survey, Applications, and Research Challenges,” 2021.
[7] Tsipouras, “Machine Learning in Biomedicine and Healthcare”, 2018.
[8] Carlos Ordonez,” A Survey of Machine Learning for Big Data Processing”, ACM Computing Surveys, 2019.
[9] Zichao Yang, Diyi Yang, Chris Dyer, “Multi-Label Text Classification with Long Short-Term Memory Recurrent Neural Network”, arXiv, 2016.
[10] Srikanta Pal and Amit Awekar, “A Survey on Text Classification Algorithms”, Artificial Intelligence Review, 2019.
[11] Karol Kurach, Krzysztof Pawłowski, Łukasz Romaszko, Marcin Tatjewski, Andrzej Janusz & Hung Son Nguyen, “Multi-label Classification of Biomedical Articles”2007.
[12] Kevin Thomas, Rohan Paul, Mia Kanzawa, “PubMeSH: Extreme Multi-label Classification of Biomedical Research,” 2019.
[13] Jacob Junior AFL, do Carmo FA, de Santana AL, Santana EEC, Lobato FMF. EvoImp: Multiple Imputation of Multi-label Classification data with a genetic algorithm. PLoS One. 2024 Jan 19;19(1):e0297147. doi: 10.1371/journal.pone.0297147. PMID: 38241256; PMCID: PMC10798481.

[14] Peskin AP, Dima AA. Classification of Journal Articles in a Search for New Experimental Thermophysical Property Data: A Case Study. Integr Mater Manuf Innov. 2017;6(2):187–96. doi: 10.1007/s40192-017-0096-1. PMID: 30984514; PMCID: PMC6459198.

[15] Rivest M, Vignola-Gagné E, Archambault É. Article-level classification of scientific publications: A comparison of deep learning, direct citation and bibliographic coupling. PLoS One. 2021 May 11;16(5):e0251493. doi: 10.1371/journal.pone.0251493. PMID: 33974653; PMCID: PMC8112690.

[16] Ishankulov T, Danilov G, Kotik K, Orlov Y, Shifrin M, Potapov A. The Classification of Scientific Abstracts Using Text Statistical Features. Stud Health Technol Inform. 2022 Jun 6;290:263-267. doi: 10.3233/SHTI220075. PMID: 35673014.

[17] Rabby G, Berka P. Multi-class classification of COVID-19 documents using machine learning algorithms. J Intell Inf Syst. 2023;60(2):571-591. doi: 10.1007/s10844-022-00768-8. Epub 2022 Nov 29. PMID: 36465147; PMCID: PMC9707112.

[18] Dernoncourt, Franck, and Ji Young Lee. "PubMed 200k RCT: A Dataset for Sequential Sentence Classification in Medical Abstracts." *arXiv.Org*, 17 Oct. 2017, arxiv.org/abs/1710.06071.

[19] Mustafa G, Usman M, Yu L, Afzal MT, Sulaiman M, Shahid A. Multi-label classification of research articles using Word2Vec and identification of similarity threshold. Sci Rep. 2021 Nov 9;11(1):21900. doi: 10.1038/s41598-021-01460-7. PMID: 34754057; PMCID: PMC8578475.

[20] Gérardin C, Wajsbürt P, Vaillant P, Bellamine A, Carrat F, Tannier X. Multilabel classification of medical concepts for patient clinical profile identification. Artif Intell Med. 2022 Jun;128:102311. doi: 10.1016/j.artmed.2022.102311. Epub 2022 Apr 26. PMID: 35534148.

[21] Wang D, Lian J, Jiao W. Multi-label classification of retinal disease via a novel vision transformer model. Front Neurosci. 2024 Jan 8;17:1290803. doi: 10.3389/fnins.2023.1290803. PMID: 38260025; PMCID: PMC10800810.

[22] Li C, Sun L, Peng D, Subramani S, Nicolas SC. A multi-label classification system for anomaly classification in electrocardiogram. Health Inf Sci Syst. 2022 Aug 25;10(1):19. doi: 10.1007/s13755-022-00192-w. PMID: 36032778; PMCID: PMC9411383.

[23] W. Liu, H. Wang, X. Shen and I. W. Tsang, "The Emerging Trends of Multi-Label Learning," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 11, pp. 7955-7974, 1 Nov. 2022, doi: 10.1109/TPAMI.2021.3119334.

[24] Zhang Y, Li X, Liu Y, Li A, Yang X, Tang X. A Multilabel Text Classifier of Cancer Literature at the Publication Level: Methods Study of Medical Text Classification. JMIR Med Inform. 2023 Oct 5;11:e44892. doi: 10.2196/44892. PMID: 37796584; PMCID: PMC10587805.

[25] Rivest M, Vignola-Gagné E, Archambault É. Article-level classification of scientific publications: A comparison of deep learning, direct citation and bibliographic coupling. PLoS One. 2021 May 11;16(5):e0251493. doi: 10.1371/journal.pone.0251493. PMID: 33974653; PMCID: PMC8112690.