Implementing NLP Algorithms for Tag Prediction on Medical Research Papers

1st Moaaz Tameer Islam
SEECS)

NUST
Islamabad, Pakistan
mtislam.bese19seecs@seecs.edu.pk

2nd Omar Ahmed
SEECS
NUST
Islamabad, Pakistan
oahmed.bese19seecs@seecs.edu.pk

***Abstract*—The lack of consistent and accurate tagging on older medical research papers is a significant challenge in searching and accessing relevant information. In this research project, we propose to address this issue by developing a Natural Language Processing (NLP) solution that can predict tags for medical research papers on Pubmed.org that currently lack them. Our approach involves training an NLP model on the title and abstract of tagged papers to predict relevant tags for untagged papers. We plan to evaluate our model using precision, recall, and F1-score metrics and explore the potential impact of our solution on enhancing the accessibility and impact of medical research.**

***Index Terms*—NLP, BERT**

I. Introduction

The accessibility and impact of medical research heavily rely on effective search and categorization systems, making tagging an essential aspect of medical research publications. However, older medical research papers may lack consistent tagging, making them difficult to search and find. This limitation can hinder research efforts and prevent the discovery of important information. To address this issue, we propose a research project that aims to develop an NLP solution that predicts tags for medical research papers on Pubmed.org that currently lack them. We plan to train our model on the title and abstract of tagged papers using state-of-the-art NLP algorithms such as BERT encoders, and other techniques that may improve our model's accuracy. We will evaluate the performance of our model using precision, recall, and F1-score metrics. The ultimate goal of this research project is to enhance the accessibility and impact of older medical research papers, thereby enabling researchers, medical professionals, and policymakers to stay up-to-date with the latest research. We expect that our solution will have broader applicability beyond medical research papers, such as scientific research publications.

II. Overall Aims and Achievable Targets

The primary goal of this research project is to develop an NLP solution that can predict relevant tags for medical research papers that currently lack them, thereby enhancing the accessibility and impact of older research. To achieve this, we will utilize state-of-the-art NLP algorithms such as BERT encoders and other techniques to train our model on the title and abstract of tagged papers. Our proposed solution will involve the following achievable targets:

Data Collection: We will collect medical research papers from Pubmed.org that lack consistent tagging.

Data Preprocessing: We will preprocess the collected data by cleaning, tokenizing, and transforming the text data into a format suitable for NLP algorithms.

NLP Model Development: We will develop an NLP model that can predict relevant tags for untagged medical research papers. Our model will be trained on the title and abstract of tagged papers using state-of-the-art NLP algorithms such as BERT encoders and other techniques to enhance its accuracy.

Model Evaluation: We will evaluate the performance of our NLP model using precision, recall, and F1-score metrics.

Model Deployment: We will deploy our NLP model to predict tags for untagged medical research papers on Pubmed.org, thereby enhancing the accessibility and impact of older research.

In summary, this research project aims to solve the challenge of inconsistent and inaccurate tagging of older medical research papers by developing an NLP solution that can predict relevant tags for untagged papers. We plan to achieve this by utilizing state-of-the-art NLP algorithms such as BERT encoders and other techniques to develop and evaluate our NLP model. The potential impact of this project includes enhancing the accessibility and impact of older medical research, enabling researchers, medical professionals, and policymakers to stay up-to-date with the latest research.

III. Dataset and Evaluation Criteria

For this research project, we will utilize medical research papers collected from Pubmed.org that lack consistent tagging. We will also explore additional data sources such as Kaggle to augment our dataset. Our dataset will be preprocessed to

ensure that it is in a suitable format for NLP algorithms, and we will randomly split it into training, validation, and test sets.

To evaluate the performance of our proposed NLP solution, we will use standard classification metrics such as precision, recall, and F1-score. We will also explore other performance metrics such as accuracy and area under the receiver operating characteristic curve (AUC-ROC). Our primary objective will be to optimize the F1-score, which is a harmonic mean of precision and recall, as it provides a balanced evaluation of our model's accuracy.

To evaluate our model's generalization capability, we will perform a k-fold cross-validation, where we will divide the dataset into k subsets and train our model on k-1 subsets, validating its performance on the remaining subset. This process will be repeated k times to ensure that we obtain a robust evaluation of our model's performance.

In summary, we will utilize medical research papers collected from Pubmed.org and additional data sources such as Kaggle to develop and evaluate our NLP solution. We will use standard classification metrics such as precision, recall, and F1-score to evaluate our model's performance and optimize its accuracy. Additionally, we will perform a k-fold cross-validation to ensure that our model's performance is robust and generalizable to new datasets.

Potential application areas where the proposed solution might help cast an impact.

The proposed NLP solution for tagging medical research papers has several potential application areas in the medical field. One of the primary applications is improving the efficiency of information retrieval in medical research. By accurately tagging research papers, our proposed solution can help researchers and medical professionals to easily find and retrieve relevant information, ultimately leading to more informed decision-making and improved patient outcomes.

Another potential application is in the development of medical knowledge graphs. Knowledge graphs are a powerful tool that can be used to represent complex relationships between medical concepts, allowing researchers to better understand the underlying mechanisms of diseases and treatments. By accurately tagging research papers, our proposed solution can help to populate and enrich these knowledge graphs, facilitating new discoveries and advancing medical research.

In addition, our proposed solution can be used to identify research gaps and areas of interest. By analyzing the tags assigned to research papers, researchers can identify areas where research is lacking and focus their efforts on these areas. This can ultimately lead to the development of new treatments and therapies for diseases and conditions that are currently under-researched.

Overall, the proposed NLP solution for tagging medical research papers has several potential application areas in the medical field, including improving information retrieval, developing medical knowledge graphs, and identifying research gaps and areas of interest. These applications have the potential to impact the medical field by facilitating new discoveries, improving decision-making, and ultimately leading to improved patient outcomes.

CEP Attribute Mapping

WP1: The proposed project requires a deep understanding of NLP algorithms, techniques, and technologies such as BERT encoders. The solution needs to be substantial enough to cover all related aspects of the problem, including data preprocessing, feature extraction, model selection, and evaluation.

WP2: The proposed project incorporates diverse scenarios and features where decisions need to be taken to select the most suitable method or model for the situation. For example, the project needs to consider the trade-offs between accuracy and interpretability when selecting the NLP algorithm.

WP7: The proposed project is designed in a modular form where the problem is divided into sub-tasks, which could be independently implemented and then integrated to form the final product. For example, the project needs to include subtasks such as data scraping, pre-processing, feature extraction, model training, and evaluation.

WP8: The proposed project is justified as a learning exercise, which provides an opportunity to gain practical experience in NLP algorithms and technologies. The project aims to develop a solution to a real-world problem and has the potential to impact the medical field by facilitating new discoveries, improving decision-making, and ultimately leading to improved patient outcomes.