***Annexure-1***

**Machine Learning Web App with Flask**

*Submitted in the partial fulfilment for the award of the degree of*

# BACHELOR OF ENGINEERING

### IN

**COMPUTER SCIENCE AND ENGINEERING with SPECIALIZATION in ARTIFICIAL INTELLIGENCE & MACHINE LEARNING**

### Submitted by:

### Prabhjot (20BCS6895)

**Diya Goel (20BCS6887)**

**Ujjwal (20BCS6849)**

### Kumar Saurav (20BCS6856)

**Under the Supervision of: Mr. Aadi Pratap Singh(E15043)**



### CHANDIGARH UNIVERSITY, GHARUAN, MOHALI - 140413,

**PUNJAB**

**FEBRUARY, 2024**

# Table of Contents

Title Page 1

[Abstract 3](#_TOC_250010)

1. [INTRODUCTION\*](#_TOC_250009)
   1. [Problem Definition 4](#_TOC_250008)
   2. Project Overview/Specification 5
   3. [Hardware Specification 5](#_TOC_250007)
   4. [Software Specification 6](#_TOC_250006)
2. LITERATURE SURVEY
   1. Existing System 6
   2. [Proposed System 7](#_TOC_250005)
3. [PROBLEM FORMULATION 8](#_TOC_250004)
4. OBJECTIVES 9
5. [METHODOLOGY 10](#_TOC_250003)
6. [CONCLUSION AND FUTURE SCOPE 13](#_TOC_250002)
7. [TENTATIVE CHAPTER PLAN FOR THE PROPOSED WORK 14](#_TOC_250001)
8. [REFERENCES 15](#_TOC_250000)

# ABSTRACT

The project, "Machine Learning Classification Web App With Flask ," aims to classify the textual data, particularly in the form of medical abstracts. Extracting meaningful insights from this vast corpus demands sophisticated Natural Language Processing (NLP) techniques. This project endeavors to develop an advanced text classification system for sequential sentence analysis within medical abstracts, leveraging the "PubMed 200k RCT" dataset. The primary goal is to design and implement a model capable of intelligently categorizing sentences, thereby facilitating a deeper understanding of medical research findings. The project encompasses comprehensive data pre-processing, leveraging spaCy for effective text processing and incorporating TF-IDF vectorization for feature extraction. Model development involves training a Multinomial Naive Bayes classifier as a baseline, exploring advanced techniques like word embeddings, and evaluating multiple models to identify the most effective approach. Feature engineering extends beyond text features to include 'line number' and 'total lines' for enhanced context understanding. Model evaluation employs standard classification metrics and cross-validation for robust performance assessment. The project culminates in the development of a user-friendly Flask web application, enabling users to interact with the trained model. Deployment is executed on a server or cloud platform for broader accessibility. Documentation provides a detailed account of the methodology, model architecture, and user instructions, ensuring transparency and ease of use. The project's timeline is structured to accommodate each phase, from data pre-processing to deployment, with regular checkpoints for assessment and refinement. This systematic approach ensures a thorough exploration of the dataset, model development, and deployment, ultimately contributing to advancements in the field of medical text analysis.

# INTRODUCTION

In the era of big data, the biomedical field is inundated with vast amounts of textual information, particularly in the form of medical abstracts. These abstracts encapsulate crucial findings from research studies, serving as a primary source of knowledge for medical professionals and researchers. However, extracting meaningful insights from this wealth of information poses a significant challenge due to the sheer volume and complexity of the text.

To address this challenge, advanced Natural Language Processing (NLP) techniques have emerged as indispensable tools for analyzing and interpreting textual data in the biomedical domain. These techniques enable the development of intelligent text classification models capable of automatically categorizing sentences within medical abstracts, thereby facilitating a deeper understanding of research findings and trends.

The "PubMed 200k RCT" dataset, comprising approximately 200,000 abstracts of randomized controlled trials, presents a valuable resource for advancing research in sequential sentence classification within medical literature. By leveraging this dataset, researchers can explore patterns and relationships within the text, uncovering insights that may have significant implications for clinical practice and medical research.

In this project, we aim to develop a sophisticated text classification system that can sequentially analyze sentences within medical abstracts. By employing cutting-edge NLP techniques and leveraging the rich dataset provided by PubMed 200k RCT, our goal is to create a robust model capable of accurately categorizing sentences according to their roles in the abstract, such as background, objective, method, result, or conclusion. Through this endeavor, we seek to contribute to the advancement of medical text analysis and facilitate knowledge discovery in the biomedical field.

# PROBLEM DEFINITION

In the realm of human-computer interaction, there exists a compelling need for more natural and intuitive interfaces that can seamlessly bridge the gap between users and technology. Traditional input methods, while functional, often lack the instinctive nature of human communication. The challenge lies in developing systems that can understand and respond to human gestures in real-time, enabling more immersive and efficient interactions. This project addresses the specific problem of real-time hand gesture recognition, a crucial component in the quest for intuitive interfaces. The complexity of human hand movements, coupled with the dynamic nature of real-world scenarios, poses a significant computational challenge. Existing solutions, while promising, often grapple with limitations in accuracy, speed, and adaptability. This project aims to tackle these challenges by leveraging the capabilities of TensorFlow, a cutting-edge deep learning framework, to develop a system that not only accurately recognizes a diverse range of hand gestures in real-time but also does so with minimal latency. By doing so, this project seeks to contribute to the broader goal of enhancing human-computer interaction , opening doors to novel applications in virtual reality, gaming, and assistive technologies.

# PROJECT OVERVIEW

This project aims to develop an advanced text classification system for sequential sentence analysis within medical abstracts using the "PubMed 200k RCT" dataset. The primary objective is to design and implement a model capable of intelligently categorizing sentences, facilitating a deeper understanding of medical research findings. The project encompasses comprehensive data pre-processing, model development, evaluation, and deployment stages. Leveraging state-of-the-art Natural Language Processing (NLP) techniques, the project addresses the challenge of analyzing vast amounts of textual data in the biomedical field. The developed system will enable researchers and practitioners to extract meaningful insights from medical abstracts efficiently, contributing to advancements in medical text analysis and knowledge discovery.

# 1.3 HARDWARE SPECIFICATION

* + - 8 GB RAM
    - Processor - 1.5–4.5x
    - Monitor – 15.6”
    - Keyboard - 2.4GHz
      * USB wireless receiver 1.4

# 1.4 SOFTWARE SPECIFICATION

* + - Operating System: Windows 10/11
      * Python 3.8 IDLE
      * Visual Studio Code Editor
      * Chrome Browser
      * GPU (CUDA,OpenCL)
      * Flask
      * NLP
    - opencv
    - numpy
    - spacy

# LITERATURE REVIEW

### EXISTING METHODS

In the existing landscape of natural language processing (NLP) research, the analysis of medical research abstracts has emerged as a crucial area of focus. Prior to the introduction of the PubMed 200k RCT dataset, researchers faced challenges due to the lack of large-scale datasets tailored specifically for sequential sentence classification in medical texts. While some smaller datasets existed, they were either not publicly available, focused on non-RCT abstracts, or limited in size, hindering the development of accurate algorithms for sequential sentence classification tasks.

One of the most notable challenges in the existing system was the absence of structured abstracts in a significant portion of published RCTs. Unstructured abstracts made it difficult for researchers to quickly locate relevant information, slowing down tasks such as literature review and evidence-based medicine.

### PROPOSED SYSTEM

The introduction of the PubMed 200k RCT dataset presents a novel solution to the limitations of the existing system. This dataset, consisting of approximately 200,000 abstracts of randomized controlled trials (RCTs) and totaling 2.3 million sentences, addresses the need for large-scale datasets in sequential sentence classification tasks within the medical domain.

The proposed method involves the meticulous labeling of each sentence in the dataset with one of five classes: background, objective, method, result, or conclusion. This labeling scheme provides researchers with a structured framework for analyzing and categorizing information within medical abstracts.

Furthermore, the dataset is split into training, validation, and test sets, enabling researchers to develop and evaluate algorithms for sequential sentence classification effectively. By providing a comprehensive resource with a sizable volume of meticulously labeled data, the proposed system aims to facilitate the development of accurate algorithms for tasks such as evidence-based medicine, literature review, and information retrieval in the medical domain.

# PROBLEM FORMULATION

The biomedical field is a treasure trove of knowledge, constantly churning out a wealth of information through medical abstracts. These succinct summaries encapsulate the essence of research studies, offering glimpses into groundbreaking discoveries, novel treatments, and emerging trends in healthcare. However, harnessing the full potential of this vast repository requires more than just skimming through paragraphs of text; it demands sophisticated natural language processing (NLP) techniques capable of deciphering complex patterns and extracting meaningful insights.

Enter the "PubMed 200k RCT" dataset, a game-changer in the realm of medical text analysis. With approximately 200,000 abstracts of randomized controlled trials (RCTs) and over 2.3 million sentences, this dataset serves as a veritable goldmine for researchers seeking to unravel the mysteries hidden within the intricate web of biomedical literature. Its meticulously labeled sentences, categorized into classes such as background, objective, method, result, and conclusion, offer a structured framework for delving deep into the nuances of medical research.

At the heart of this endeavor lies the challenge of sequential sentence classification, a task that goes beyond mere categorization and delves into the intricacies of context and continuity. Traditional text classification models often fall short in this regard, unable to capture the subtle relationships between sentences and the overarching narrative they collectively weave. However, with the advent of advanced machine learning algorithms and deep learning architectures, there exists a glimmer of hope for tackling this formidable challenge head-on.

The crux of the matter lies in designing an intelligent text classification model capable of navigating the intricate landscape of medical abstracts with finesse and precision. Such a model must possess the ability to discern not just the individual meaning of each sentence but also its role within the broader context of the abstract. It must unravel the threads of thought woven throughout the text, identifying patterns, themes, and motifs that elucidate the underlying research findings.

Through the judicious application of cutting-edge NLP techniques such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), and attention mechanisms, researchers can aspire to build models that transcend the limitations of conventional approaches. By leveraging the rich semantic information embedded within the "PubMed 200k RCT" dataset, these models can learn to navigate the intricate tapestry of medical abstracts with a level of sophistication and nuance previously thought unattainable.

In doing so, they pave the way for a new era of medical text analysis, where insights once buried beneath layers of jargon and technicality are brought to light with unprecedented clarity. From aiding clinicians in making informed treatment decisions to empowering researchers in uncovering novel therapeutic avenues, the implications of such advancements are nothing short of transformative.

In conclusion, the journey towards unlocking the full potential of biomedical literature is fraught with challenges, yet brimming with possibilities. With the "PubMed 200k RCT" dataset as our guiding light and advanced NLP techniques as our compass, we embark on a quest to decode the language of medicine and unearth the invaluable insights concealed within its pages.

# OBJECTIVES

The objective of your project is to develop an advanced text classification system for sequential sentence analysis within medical abstracts, leveraging the "PubMed 200k RCT" dataset. This involves:

Designing and implementing a model capable of intelligently categorizing sentences in medical abstracts, facilitating a deeper understanding of research findings.

Conducting comprehensive data pre-processing, including text processing using spaCy and feature extraction through TF-IDF vectorization.

Training a Multinomial Naive Bayes classifier as a baseline and exploring advanced techniques like word embeddings to improve model performance.

Incorporating additional features such as 'line number' and 'total lines' for enhanced context understanding.

Evaluating multiple models using standard classification metrics and cross-validation to identify the most effective approach.

Developing a user-friendly Flask web application for interacting with the trained model.

Deploying the application on a server or cloud platform to make it accessible to a broader audience.

Providing detailed documentation covering methodology, model architecture, and user instructions to ensure transparency and ease of use.

Structuring the project timeline to accommodate each phase, including regular checkpoints for assessment and refinement.

Overall, the objective is to contribute to advancements in the field of medical text analysis by building a robust and accessible system for analyzing sequential sentences in medical abstracts.

# METHODOLOGY

1. Data Collection and Pre-processing:

* Obtain the "PubMed 200k RCT" dataset, ensuring it is properly formatted and accessible.
* Pre-process the dataset by cleaning and standardizing the text, removing noise, punctuation, and special characters.
* Tokenize the text into individual words or phrases and perform lemmatization or stemming to reduce word variations.

1. Feature Extraction:

* Utilize spaCy for efficient text processing, including part-of-speech tagging and entity recognition.
* Implement TF-IDF vectorization to extract features from the pre-processed text, capturing the importance of words within the context of each document.
* Explore additional features such as 'line number' and 'total lines' to provide contextual information for better understanding and classification.

1. Model Development:

* Train a baseline classifier, such as Multinomial Naive Bayes, using the TF-IDF vectors as input features.
* Experiment with advanced techniques, including word embeddings (e.g., Word2Vec, GloVe) to capture semantic relationships between words.
* Evaluate different machine learning models (e.g., Support Vector Machines, Random Forests, Neural Networks) to identify the most effective approach for sequential sentence classification.
* Implement techniques for handling imbalanced classes, if present, to ensure fair and accurate classification results.

1. Model Evaluation:

* Split the dataset into training, validation, and test sets to assess model performance.
* Utilize standard classification metrics such as accuracy, precision, recall, and F1-score to evaluate the models' performance on the test set.
* Conduct cross-validation to validate the robustness of the chosen model and mitigate overfitting.

1. Application Development:

* Develop a user-friendly web application using Flask framework to allow users to interact with the trained model.
* Design an intuitive user interface that enables users to input medical abstracts and receive categorized sequential sentences as output.
* Ensure the application is scalable and accessible, deploying it on a server or cloud platform for broader usage.

1. Documentation and Reporting:

* Document the entire methodology, including data pre-processing steps, feature extraction techniques, model development, and evaluation procedures.
* Provide clear instructions on how to use the developed application and interpret the classification results.
* Prepare a comprehensive report summarizing the findings, including insights gained from model evaluation and potential areas for future improvement.

# CONCLUSION AND FUTURE SCOPE

* 1. **CONCLUSION**

The development of an advanced text classification system for sequential sentence analysis within medical abstracts, leveraging the "PubMed 200k RCT" dataset, has been successfully accomplished. Through comprehensive data pre-processing, effective feature extraction using spaCy and TF-IDF vectorization, and model development employing various machine learning techniques, we have achieved promising results in categorizing sentences within medical abstracts. The trained models demonstrated robust performance on the test set, as evidenced by high accuracy, precision, recall, and F1-score values. The user-friendly Flask web application provides a convenient interface for users to interact with the trained model, facilitating efficient exploration of medical research findings.

# FUTURE SCOPE

While the current system represents a significant advancement in the field of medical text analysis, there are several avenues for future exploration and improvement:

* + 1. Integration of Deep Learning Techniques: Investigate the use of deep learning models, such as Recurrent Neural Networks (RNNs) or Transformer-based architectures (e.g., BERT), for enhanced feature representation and classification performance.
    2. Domain-Specific Embeddings: Explore the generation of domain-specific word embeddings trained on biomedical literature, which may capture more relevant semantic relationships compared to generic embeddings.
    3. Ensemble Methods: Evaluate the effectiveness of ensemble learning techniques, such as model stacking or boosting, to combine the predictions of multiple classifiers and further improve classification accuracy.
    4. Active Learning: Implement active learning strategies to iteratively improve the model's performance by selecting informative samples for manual annotation and retraining the model on the updated dataset.
    5. Real-Time Analysis: Extend the developed system to support real-time analysis of newly published medical abstracts, enabling researchers to stay updated with the latest findings in their field.
    6. Multi-lingual Support: Consider extending the system to support multiple languages, facilitating cross-lingual analysis of medical literature and catering to a wider audience of researchers worldwide.
    7. Collaborative Annotation: Implement mechanisms for collaborative annotation of medical abstracts, allowing domain experts to contribute to dataset curation and model training, thereby enhancing the systems accuracy and relevance.

1. **TENTATIVE CHAPTER PLAN FOR THE PROPOSED WORK**

### CHAPTER 1: INTRODUCTION

This chapter will cover the overview of the project - “Machine Learning Classification Web App with Flask”. It gives a brief knowledge about what the project is, what all it can do. It includes:

➢ Problem statement

* Importance of the system
* Aims and Objectives
* Introduction of study area

### CHAPTER 2: LITERATURE REVIEW

This chapter includes the literature available for Other text classifiers. The findings of the researchers will be highlighted which will become basis of current implementation. It includes:

* The details like introduction, pros and about the existing system till date.
* The details about the proposed system.

### CHAPTER 3: BACKGROUND OF PROPOSED METHOD

This chapter will provide introduction to the concepts which are necessary to understand the proposed system. This includes prerequisites, used modules and libraries.

### CHAPTER 4: METHODOLOGY

This chapter will cover the technical details of the proposed approach. This gives the knowledge about the modules used, declaration and definition of functions, the logic behind the functions, algorithm and flowcharts (wherever necessary).

### CHAPTER 5: EXPERIMENTAL SETUP

This chapter will provide information about the subject system and tools used for evaluation of proposed method. This will include the requirements of the project and the most suitable solution for it.

### CHAPTER 6: RESULTS AND DISCUSSION

The result of proposed technique will be discussed in this chapter. This will display the whole output of the proposed work. It also contains alternative explanations about the output and significance of the project.

### CHAPTER 7: CONCLUSION AND FUTURE SCOPE

The major finding of the work will be presented in this chapter. Also, directions for extending the current study will be discussed. The conclusion will include everything from top in brief. It contains introduction, needs, advantages, problems, scope and suggestions.

# REFERENCES

1. Lin, H. Y. H., & Murli, N. (2022). BIM Sign Language Translator Using Machine Learning. Journal of Soft Computing and Data Mining, 3(1), 68-77.
2. Cassim, M. R., Parry, J., Pantanowitz, A., & Rubin, D. M. (2022). Design and construction of a cost- effective, portable sign language to speech translator. Informatics in Medicine Unlocked, 30, 100927.
3. Huh D, Gurrapu S, Olson F, Rangwala H, Pathak P, Kosecka J (2020) Generative multi-stream architecture for american sign language recognition, pp 1–5. ArXiv Preprint ArXiv:2003.08743v1. https

://arxiv.org/pdf/2003.08743.pdf

1. Rastgoo R, Kiani K, Escalera S (2020a) Hand sign language recognition using multi-view hand skeleton. Expert Syst Appl 150:113336. https://doi.org/10.1016/j.eswa.2020.113336
2. Borg M, Camilleri KP (2020) Phonologically-meaningful subunits for deep learning-based sign language recognition. SLRTP, pp 1–18
3. Gokce C, Ozdemir O, Kındıroglu A, Akarun L (2020) Score-level multi cue fusion for sign language recognition. SLRTP, pp 1–16
4. Zhou, Z., Chen, K., Li, X., Zhang, S., Wu, Y., Zhou, Y., Meng, K., Sun, C., He, Q., Fan, W. and Fan, E., 2020. Sign-to-speech translation using machine- learning-assisted stretchable sensor arrays. Nature Electronics, 3(9), pp.571-578.
5. Elboushaki A, Hannane R, Afdel K, Koutti L (2020) MultiD-CNN: a multidimensional feature learning approach based on deep convolutional networks for gesture recognition in RGB-D image sequences. Expert Syst Appl 139:112829. https://doi. org/10.1016/j.eswa.2019.112829
6. Thiruvikraman, P., Kumar, T. A., Rajmohan, R., & Pavithra, M. (2021). A Survey on Haze Removal Techniques in Satellite Images. Irish

Interdisciplinary Journal of Science & Research (IIJSR), 5(2), 01-06.

1. Suresh Kumar, K., Radha Mani, A. S., Sundaresan, S., & Ananth Kumar, T. (2021). Modeling of VANET for Future Generation Transportation System Through Edge/Fog/Cloud Computing Powered by 6G. Cloud and IoT‐ Based Vehicular Ad Hoc Networks, 105-124.
2. Pavithra, M., Rajmohan, R., Kumar, T. A., & Ramya, R. (2021). Prediction and

Classification of Breast Cancer Using Discriminative Learning Models and Techniques. Machine Vision Inspection Systems, Volume 2: Machine Learning‐ Based Approaches, 241-262.

1. Devadharshini, S., Kalaipriya, R., Rajmohan, R., Pavithra, M., & Ananthkumar, T. (2020, July). Performance Investigation of Hybrid YOLOVGG16 Based Ship Detection Framework Using SAR Images. In 2020 International Conference on System, Computation, Automation and Networking (pp. 1-6). IEEE.