**20**

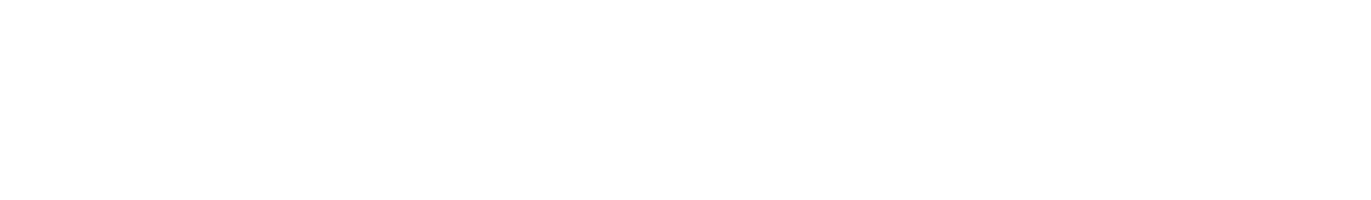
**AML**

**-**

**)**

**(B**

**6**



DEPARTMENT OF APEX INSTITUTE OF TECHNOLOGY

1. **Project Title:**

Med Chronicle: Unravelling Insights through Sequential Sentence Analysis in PubMed RCT

1. **Problem Statement:**

The biomedical field generates vast amounts of information, particularly in the form of medical abstracts. Extracting meaningful insights from these abstracts requires advanced natural language processing (NLP) techniques. The "PubMed 200k RCT" dataset presents an opportunity for sequential sentence classification in medical abstracts, aiming to uncover patterns and relationships within the text. The challenge is to design an intelligent text classification model capable of categorizing sentences sequentially, contributing to a deeper understanding of medical research findings.

1. **Project Scope and planning:**

The project aims to develop an advanced text classification system for sequential sentence analysis within medical abstracts, leveraging the "PubMed 200k RCT" dataset. The primary objective is to design and implement a model that can intelligently categorize sentences, allowing for a deeper understanding of medical research findings. The scope encompasses comprehensive data pre-processing, utilizing spaCy for effective text processing, and incorporating TF-IDF vectorization for feature extraction. Model development will involve training a Multinomial Naive Bayes classifier as a baseline, exploring advanced techniques such as word embeddings, and evaluating multiple models to identify the most effective approach. Feature engineering will extend beyond text features to include 'line number' and 'total lines' for enhanced context understanding. Model evaluation will employ standard classification metrics and cross-validation for robust performance assessment. The project will culminate in the development of a user-friendly Flask web application, allowing users to interact with the trained model. Deployment will be executed on a server or cloud platform for broader accessibility. Documentation will provide a detailed account of the methodology, model architecture, and user instructions, ensuring transparency and ease of use. The project timeline is structured to accommodate each phase, from data pre-processing to deployment, with regular checkpoints for assessment and refinement. This approach ensures a systematic and thorough exploration of the dataset, model development, and deployment, ultimately contributing to advancements in the field of medical text analysis.

1. **Task definition:**

The project tasks include data pre-processing using spaCy for tokenization and lemmatization, and TF-IDF vectorization for feature extraction. Model development involves training a baseline Multinomial Naive Bayes classifier, exploring word embeddings, and evaluating models through metrics and cross-validation. Feature engineering tasks include incorporating 'line number' and 'total lines' for contextual understanding. Additionally, a user-friendly Flask web app will be created for model interaction, deployed on a chosen platform. Documentation will detail the methodology, model architecture, and deployment process for user clarity.

1. **Requirements:**

* NumPy
* Pandas
* Spacy
* Sklearn
* TF-IDF Vectorizer

# STUDENTS DETAILS

|  |  |  |
| --- | --- | --- |
| **Name** | **UID** | **Signature** |
| Prabhjot Singh | 20BCS6897 |  |
| Diya Goel | 20BCS6887 |  |
| Ujjwal | 20BCS6849 |  |
| Kumar Saurav | 20BCS6856 |  |

# APPROVAL AND AUTHORITY TO PROCEED

We approve the project as described above, and authorize the team to proceed.

|  |  |  |
| --- | --- | --- |
| **Name** | **Title** | **Signature (with Date)** |
| Aadi Partap Singh (E15043) | Supervisor |  |
| Malti Rani(E14816) | Evaluation Panellist’s |  |
| Sheetal Laroiya(E15433) | Evaluation Panellist’s |  |