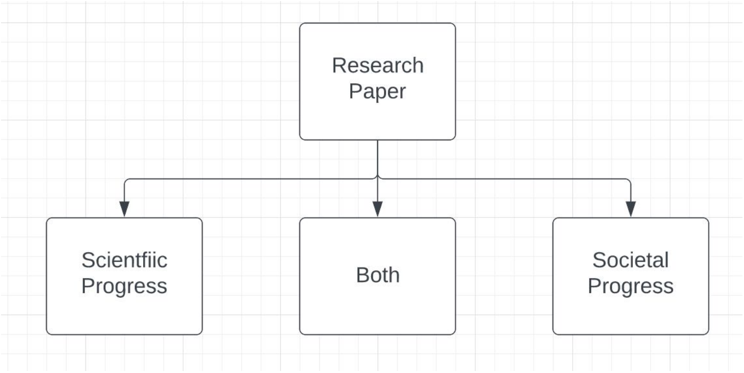
Classification of research papers

Research Problem, Questions & Objectives

Problem and Aims:

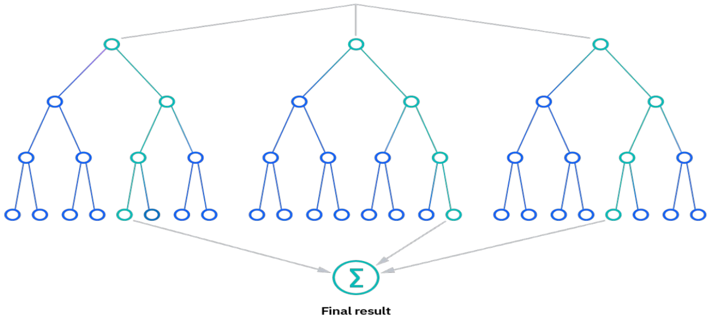
- Develop a multi-class classification model for labelling scientific research papers.
- Classifier should both be accurate and efficient



Research Problem, Questions & Objectives

Research questions and objectives:

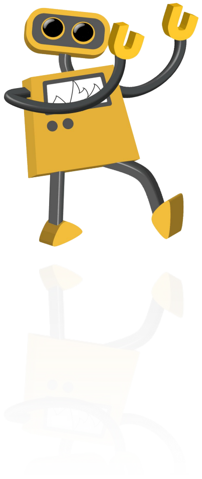
- Does SciBERT perform better as part of the pre-processing stage?
- Will deep learning models outperform traditional baseline methods



Literature Review

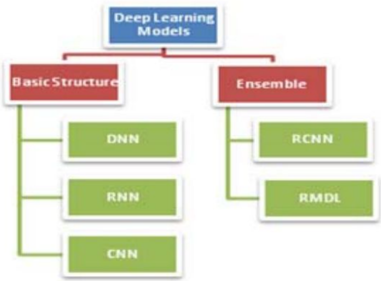
Performance Evaluation of Deep Learning Algorithms in Biomedical Document Classification

- ❑ Classified biomedical documents
- ❑ Used three datasets
- ❑ ML Techniques
  SVM / PPN / SGD / PA / Ridge



Research Gaps

- ❑ Pre trained word embedding
- ❑ Achieving better accuracy in classification



$$Accuracy = \frac{\sum_{i=1}^{m} \frac{tp_i + tn_i}{tp_i + fp_i + fn_i + tn_i}}{m}$$

$$\Pr ecision_{Weighted} = \frac{\sum_{i=1}^{m} \left| y_i \right| \frac{tp_i}{tp_i + fp_i}}{\sum_i^m \left| y_i \right|}$$

$$\operatorname{Re} call_{Weighted} = \frac{\sum_{i=1}^{m} \left| y_i \right| \frac{tp_i}{tp_i + fn_i}}{\sum_i^m \left| y_i \right|}$$

$$F1 - Score_{Weighted} = \frac{\sum_1^m \left| y_i \right| \frac{2tp_i}{2tp_i + fp_i + fn_i}}{\sum_i^m \left| y_i \right|}$$

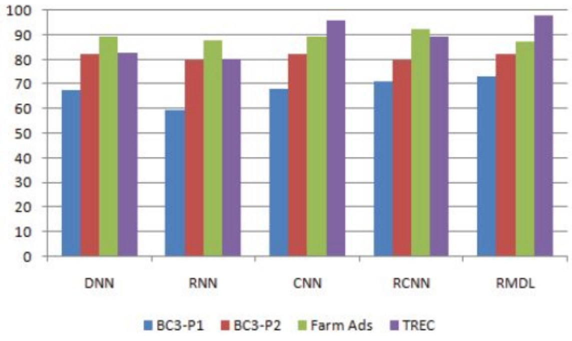


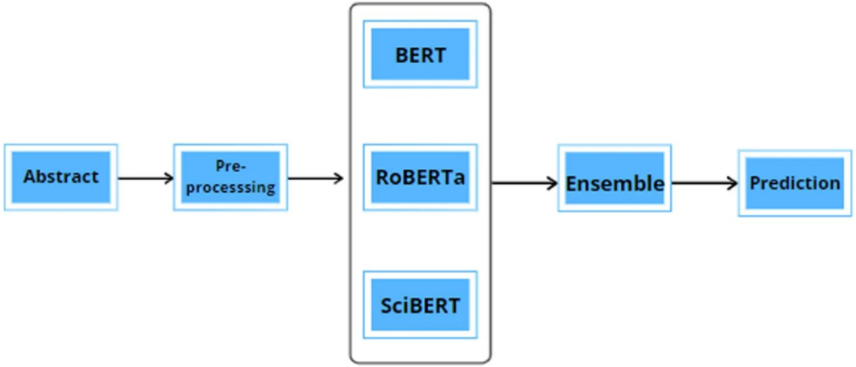
Fig.2.Comparision of Classification Accuracy (%)

Literature Review

Domain Identification of Scientific Articles Using Transfer Learning and Ensembles

★ Used 35,000 scientific article abstracts

★ Compared BERT, SciBERT and RoBERTa models accuracy



Research Gaps

★ Relies on a large amount of labelled data for pre-training

★ Only considers the textual content of scientific articles.

★ Other modalities, such as figures or tables, are not considered.

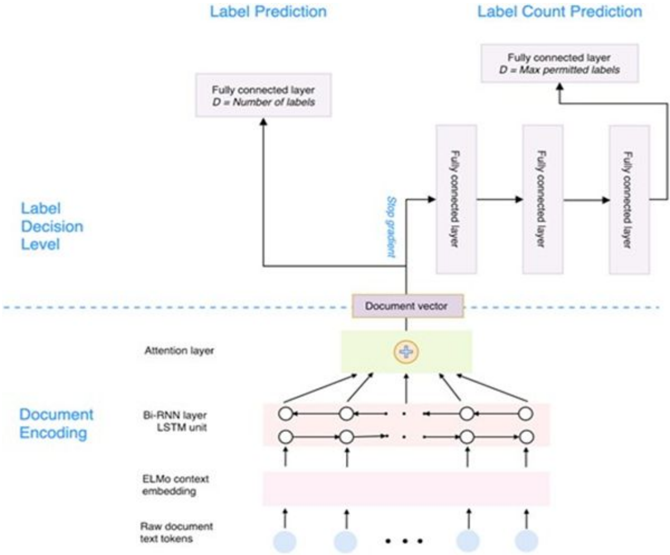
Literature Review

ML-Net: multi-label classification of biomedical texts with deep neural networks

- Classifies medical documents
- Uses ELMo word embeddings and RNN
- End-to-end deep learning framework

Research Gaps

- Hard to classify hierarchical data
- Incorrect predictions when met with skewed datasets



Research Gaps

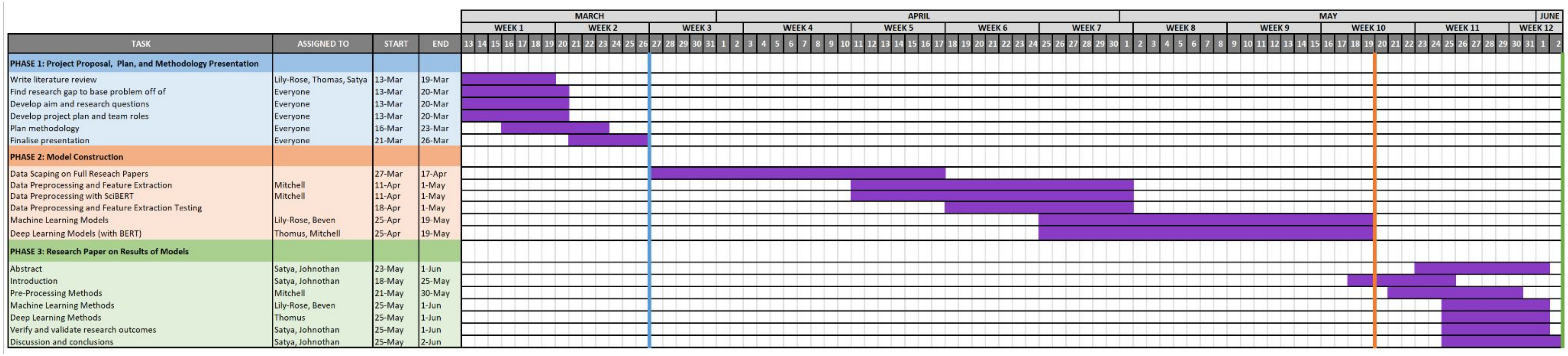
- Limited research on multi-class classification of scientific documents
- Lack of consensus on the most suitable methods for scientific document classification
- Unclear tradeoff between deep learning classifiers and traditional baselines
- Limited research on BERT and SCIBERT integration as classification for scientific documents

Project Plan - Team Roles

Mitchell	Thomas	Lily	Satya	Bevan	Jonothan
Group Leader	CNN	Random forest	Report Writer	SVM model	Report Writer
Pre-processing	RCNN				
Paper scraper					
SciBert					

Project Plan

Milestone 2 - Phase 2 Completed

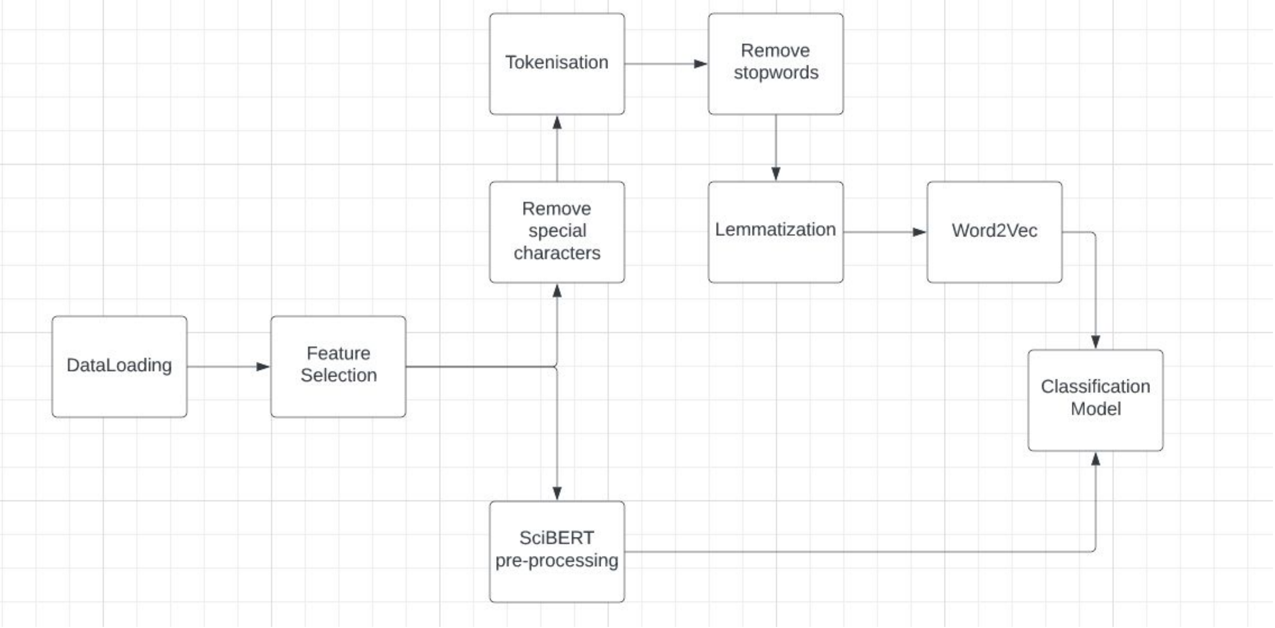


TASK	ASSIGNED TO	START	END
PHASE 1: Project Proposal, Plan, and Methodology Presentation			
Write literature review	Lily-Rose, Thomas, Satya	13-Mar	19-Mar
Find research gap to base problem off of	Everyone	13-Mar	20-Mar
Develop aim and research questions	Everyone	13-Mar	20-Mar
Develop project plan and team roles	Everyone	13-Mar	20-Mar
Plan methodology	Everyone	16-Mar	23-Mar
Finalise presentation	Everyone	21-Mar	26-Mar
PHASE 2: Model Construction			
Data Scaping on Full Reseach Papers		27-Mar	17-Apr
Data Preprocessing and Feature Extraction	Mitchell	11-Apr	1-May
Data Preprocessing with SciBERT	Mitchell	11-Apr	1-May
Data Preprocessing and Feature Extraction Testing		18-Apr	1-May
Machine Learning Models	Lily-Rose, Beven	25-Apr	19-May
Deep Learning Models (with BERT)	Thomus, Mitchell	25-Apr	19-May
PHASE 3: Research Paper on Results of Models			
Abstract	Satya, Johnothan	23-May	1-Jun
Introduction	Satya, Johnothan	18-May	25-May
Pre-Processing Methods	Mitchell	21-May	30-May
Machine Learning Methods	Lily-Rose, Beven	25-May	1-Jun
Deep Learning Methods	Thomus	25-May	1-Jun
Verify and validate research outcomes	Satya, Johnothan	25-May	1-Jun
Discussion and conclusions	Satya, Johnothan	25-May	2-Jun

Milestone 1 - Phase 1 Completed

Milestone 3 - Phase 3 Completed

Project Methodologies - Preprocessing



Project Methodologies - ML Model SVM

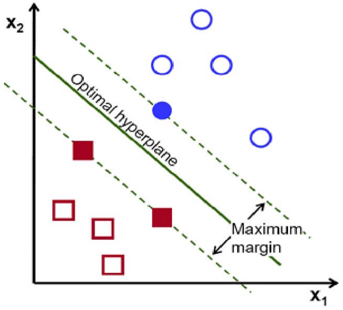
Algorithm that determines the best decision boundary between vectors belonging to a given group and those not

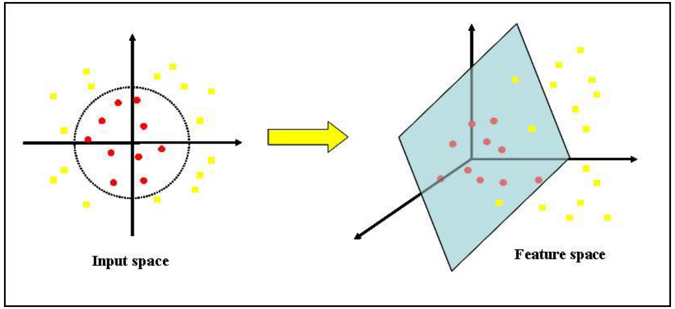
How it works:

- ❑ Best decision boundary between vectors
- ❑ Optimal hyperplane that separates the different classes

Advantages:

- ❑ Handles high-dimensional
- ❑ High volume of dataset datasets with many different features
- ❑ Keeps high accuracy





Project Methodologies - ML Model Random Forest

Algorithm that combines multiple decision trees to create a more accurate and stable model

How it works:

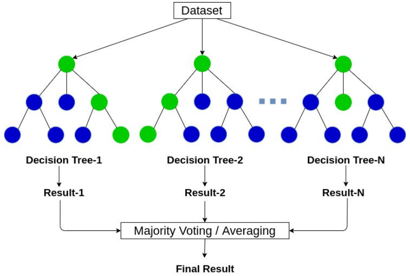
★ Combining many decision trees to make a final result

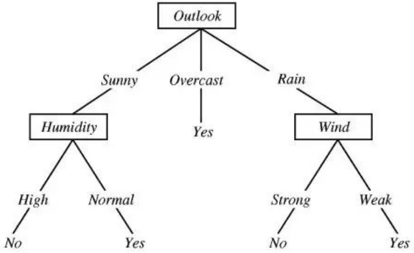
★ Each feature is splited at each node of the trees

Advantages:

★ Captures complex interactions between large number of features

★ Reduces overfitting and makes the model more accurate

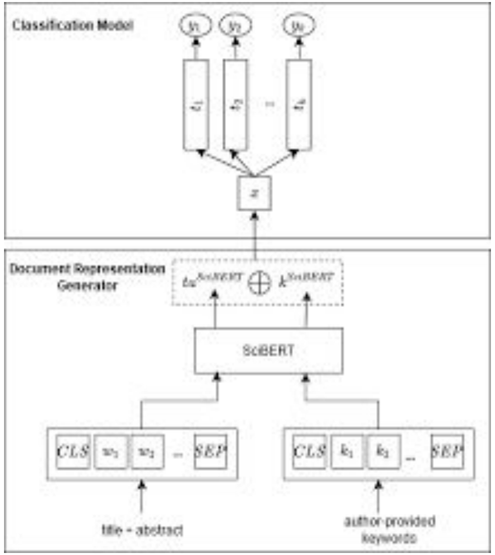




Project Methodologies - ML Model SciBERT

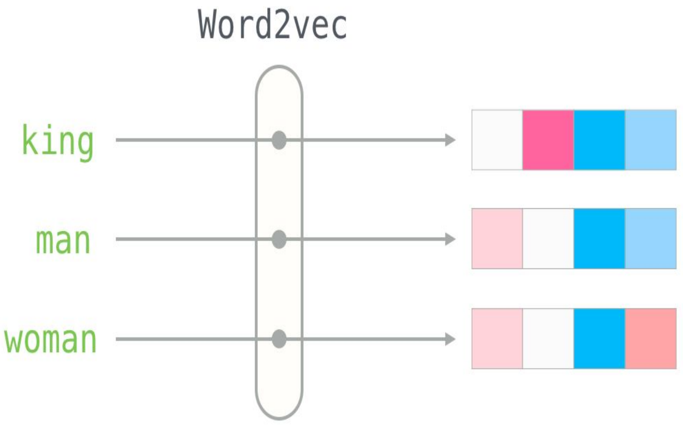
- BERT is a language model that is pre-trained on various types of text
- SciBERT is a variant of BERT specifically trained on scientific text

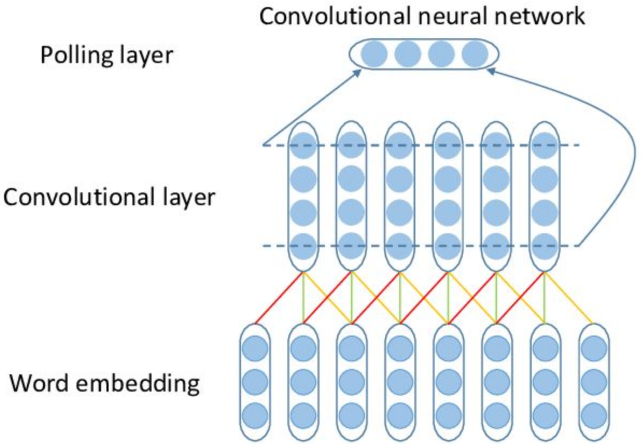
- SciBERT is the most valuable resource available for us in this project
- It is trained on not only abstracts but the entire text itself and can provide state-of-the-art results



SciBERT model example

Project Methodology - CNN

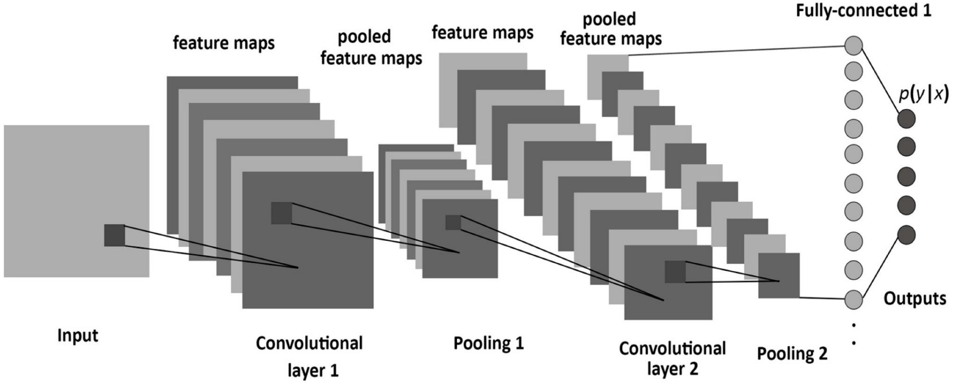




Project Methodology - Region Based Convolutional Neural Network

- Combination of Convolutional Neural Network (CNN) and recurrent neural network (RNN)

1. First divides the input into smaller regions
2. The region goes through the convolutional layer to extract features
3. Fully connected layer classifies depending on the features



Project Methodology - Evaluation of the models

- SVM
- Random Forest
- CNN
- RCNN
- SCIBERT

SVM and SciBERT will be theoretically the most effective model

Thanks for listening!