Assessment Task 3

Topic: Deep Learning-based Multi-class Classification for Scientific Documents

Abstract

The following paper presents a deep learning based approach to the multi-class classification of scientific papers. Through our research, we aim to assist individuals or organisations that want to manage their research output effectively. To accomplish this, we train SVM, Random forest, SciBERT and CNN classifiers that attempt to classify papers based on their contributions to societal progress, scientific progress or both. The results showed the best performing model was SciBert, we have also identified shortcomings of our project for future research iterations.

Introduction

In recent times, research across all fields has been exponentially growing. On a global scale, researchers generate more than 2 million papers [1] every year, creating an urgent need for appropriate data organisation practices to be in place for the efficient management of these papers. Furthermore, with the rise of artificial intelligence in the past few decades, it is evident that this task would be highly suited to be completed by an AI model, saving time, effort and financial resources. As a solution to this prevalent issue, we propose an AI based approach to the classification of scientific papers based on their contribution to societal progress, scientific progress or both. Through our research, we aim to simplify this data organisation task and create a novel classifier that can be applied to various different fields.

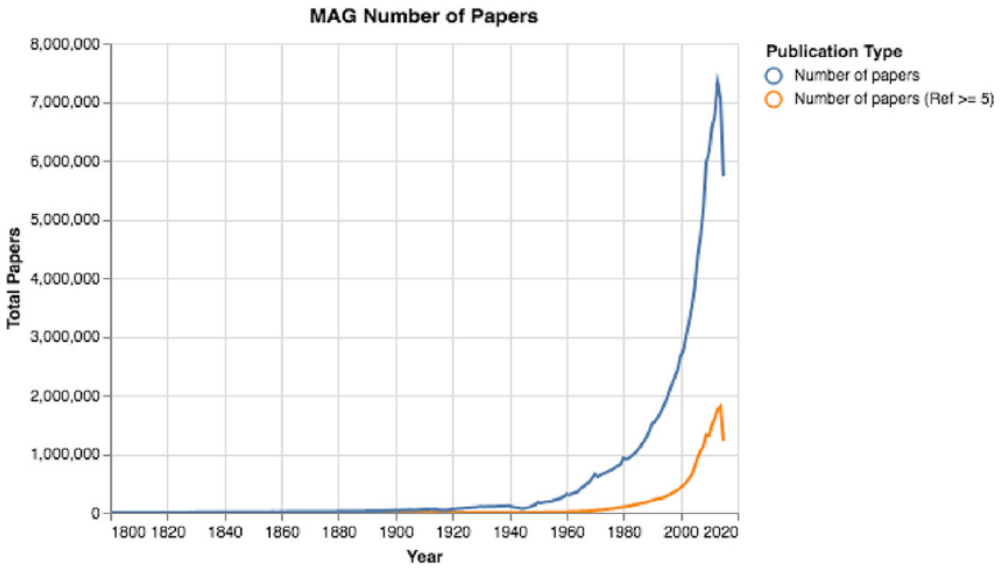


Figure 1. Volume of New Research Papers Over Time (Fire, 2019)

Literature Reviews

We conducted three comprehensive literature reviews before beginning this project. In our first literature review [2], we explored a comparison of different deep learning algorithms in the task of biomedical document classification. This literature review enabled us to understand the effectiveness of using pre-trained word embedding models like Glove, stating that these models could assist in achieving better document representation. The paper presented a comparison of deep learning models against machine learning models, of which the deep learning models were undoubtedly much more effective when given a large dataset. However, machine learning models had much better performance when given a smaller dataset.

The second literature review [3] discussed the importance of domain identification in scientific articles and its impact on search and recommendation systems. It introduces transfer learning and ensembles as techniques to improve domain identification and provides an overview of the application of transfer learning in natural language processing (NLP). The authors pre-processed the dataset by removing stop words, punctuations, and special characters, as well as applying tokenization and stemming to reduce data dimensionality and enhance model efficiency. They used BERT, RoBERTa, and SciBERT models and compared their performance. It was concluded that SciBERT performed the best, thereby informing our understanding on what model to choose. This information led to our decision to use SciBERT as a classifier and a pre-processing method for our machine and deep learning models. The paper also acknowledged potential research gaps, such as the reliance on labelled data for pre-training and the omission of other modalities like figures or tables.

Our final literature review [5], covered a deep learning framework that classified biomedical documents. The proposed network consists of a document encoding network, a label prediction network and label count prediction network, all of which work together to classify biomedical documents into various categories. The model used embeddings from language models (ELMo) which yet again proved that word embeddings were an ideal choice for such a task. As there is very limited research related to our task, this paper provided us with valuable insights regarding the best architectures associated with our project.

Review of Project proposal

In our initial project proposal, we developed our research problems, aims, questions and objectives. We also developed a project plan which detailed the roles and responsibilities of each group member and a structured methodology which involved using five models; SVM, Random Forest, SciBERT, CNN and RCNN. However, after considering the feedback that we received from the peer reviews [4] we chose to remove the RCNN model from our project. Consequently, our research questions and associated objectives have been revised to:

Question 1: Does the use of word embedding models such as SciBERT improve accuracy while being used as a classifier or as a part of the pre-processing steps.

> **Objective 1:** Use SciBERT as a pre-processing step and also as a classifier. Then compare the effectiveness of the models which use SciBERT in the pre-processing stages.

Question 2: Will the deep learning models be able to produce more accurate results than traditional baseline models currently can?

> **Objective 1:** Develop baseline ML models using SVM and random forest for classification
>
> **Objective 2:** Develop an accurate Deep Learning model using CNN models for classification

It was highly evident from our proposal that we had a lack of knowledge regarding the deep learning models as we did not provide much explanation on which specific architecture we would be using. This fact was correctly identified in our peer reviews, however, after much consideration, we chose to keep the CNN model, but decided to remove the RCNN model as it required a higher level of understanding on NLP semantics. Furthermore, considering the size of the dataset provided (1193 papers), it would be highly unlikely that a more complex deep learning model like RCNN would perform better as they often require larger datasets.

The peer review also identified that our classification into the categories of societal progress, scientific progress or both could introduce bias as these categories are not necessarily mutually exclusive. Although this is a valid point, considering the resources available to us in this project, classifying these papers into more specific categories would be highly challenging. So, we hope that this project provides a good baseline for future researchers who wish to improve our solution.

Research Design

As mentioned in the previous section, our updated methodology involves training four models; SVM, Random Forest, SciBERT and CNN. Each of these models are trained using the SciBERT pretrained model, NLTK preprocessing and a special preprocessing method for CNN's which combines the keras library with NLTK. Each phase of our experiments are detailed in the following sections.

Dataset preparation

The dataset for this task is provided in an excel sheet, with 1193 samples. Each of these samples are attached with 15 features that detail different facts about each research paper. However, to determine the impact of a particular research paper, we only require two features; the paper's title and associated abstract. As the title of the paper reveals much about a paper's content, this is an obvious choice. The abstract may also contain certain keywords that reveal the impact of a paper. Hence, these two features are obvious choices for this task. After determining our features, we read the excel sheet into our system using the pd.read_excel function available in the pandas library in Python.

Preprocessing

NLTK Preprocessing

The first step in our preprocessing stage was to remove the 13 unnecessary features from the dataset, leaving only the two required features. Then, using the natural language toolkit (NLTK) library in python we tokenise our data and remove any special characters and stopwords as they do not assist the training process [7]. Stopwords are words that have no relevance to our context and removing these words allows for a simpler model with smaller training times. We also use lemmatization, which essentially reduces a word to its root form, i.e. words like "running", "ran" and "runner" would be simplified to "run". Effectively, this process normalises the data making it easier for our model to group words together.

After preprocessing we pass the features into the word2vec word embeddings model, which converts the words into a numerical format to allow our model to capture the semantic relationship between different words. We finally also split the dataset into train, test and validation sets, with a 70-15-15 split into each of these sets.

SciBERT preprocessing

Along with NLTK, we also apply SciBERT as a preprocessing step for all our models. To implement this effectively, we use the SciBERT tokenizer to encode our data and then apply the SciBERT model on the encoded data, providing us with preprocessed data. After this stage, we split the dataset into train, test and validation sets, in the same way we did for NLTK preprocessing. As SciBERT is specifically trained to better understand and represent

scientific text (See figure 1), we theorise that this preprocessing step would be more effective for our model.

	Corpus	# of words
BERT	English Wiki	2.5B
	BooksCorpus	0.8B
SCIBERT	Biomedical	2.5B
	Computer Science	0.6B

Figure 2. Information on SciBERT

CNN preprocessing

As the structure of a CNN model contains its own word embedding, we must preprocess the data in a different way so that our CNN architecture can accept the input data. After reading the data into our system, we take in the title and abstract features and split the data into train, test and validation sets (70-15-15 split). The data is then tokenized, mapping each word to a specific number. This vector of values is what is passed into the CNN. The vector must also be padded to ensure it is the same length every time as the CNN's input is a set shape.

The CNN model using SciBERT however forgoes the word embedding layer as all of the word embedding is done by SCIBERT, as such the code is the same as used in the other models.

Models

SVM

SVM (Support Vector Machines) is a powerful classifier that aims to find the optimal hyperplane to separate different classes in the feature space while maximising the margin between them [8]. It is capable of handling both linear and non-linear classification problems by utilising kernel functions to map the data into higher-dimensional spaces. We chose SVM for paper classification due to its ability to classify effectively on small, imbalanced datasets and its potential for generalisation to unseen data. Furthermore, this model is a very popular baseline model in machine learning, and hence could be an effective model for comparison.

During the training phase, SVM learned the decision boundaries between classes based on the extracted features from the Word2Vec vectors. The objective was to find the hyperplane that best separates the classes and maximises the margin. A list of hyper parameters, their values and associated intuitions are listed in the table below:

Hyperparameter	Value (Both models)	Explanation
Kernel	RBF	The RBF kernel was used as we believed it could capture any complex decision boundaries between social and scientific papers.
C	1.0	C is the penalty parameter in SVM, which represents misclassification or error term, which can help tune the model.
random_state	42	The random seed, so that experiment can be run subsequent times to tune the model.

Table 1. SVM model hyperparameter settings

Random Forest

Random Forest is an ensemble learning model that combines multiple decision trees to make predictions [9]. Each decision tree is trained on the same subset of the training data and uses the associated features to make decisions. The final prediction is made by aggregating the predictions of individual trees. We chose random forest due to its ability to capture complex relationships with low overfitting. Similar to SVM, this model is also a popular baseline model and may be effective when provided a small dataset.

The random forest model learns from the ensemble of decision trees to make accurate predictions based on the combined ‘TI_vector’ (title vector), and ‘AB_vector’ (abstract vector) columns. A list of hyper parameters, their values and associated intuitions are listed in the table below:

Hyperparameter	Value (NLTK, SciBERT)	Explanation
n_estimators	1000, 30	As the SciBERT tokenizer already ensures that the input is pretrained (consequently a bit more complex), we have a lower value for this model and a higher value for the NLTK version.
max_depth	3, 4	As there are fewer trees in the SciBERT version, we use a higher maximum depth to ensure the model can be more accurate. In contrast, the NLTK version has many more trees, hence we have set a lower value to prevent overfitting.
min_samples_leaf	2, 3	Yet again, as we have fewer trees, to capture the finer details in the input, we use a higher value for this parameter in the SciBERT version.
n_jobs	-1, 2	To maximise the use of computational resources, we use -1 for the NLTK version and a value of 2 ensures the SciBERT version is running 2 jobs at most. This is due to the already complex nature of the input data mentioned previously.
random_state	42, 42	Ensures results can be replicated.

Table 2. Random Forest hyperparameter settings

SciBERT

The SciBERT classifier we have utilised for this task involves using transfer learning leveraging the allenai github repository [6] which contains a fully pre trained model trained on scientific text [10]. The model first maps the tokenized input into their corresponding input embeddings. These embeddings are then passed into the a transformer encoder layer, consisting of two sub layers which are:

1. Self attention mechanism - This layer captures contextual information associated with a word based on its association with other words thus allowing the model to focus on different words at the same time.
2. Feed forward neural network - This network captures any non-linear relationships, allowing the model to recognise any unusual patterns. It consists of multiple fully connected layers with different activation functions.

The input is then passed through a max pooling layer which ensures that the input representations are essentially shrunk down or summarised. Finally, a feed forward neural network is used to classify the input into its respective classes using softmax activation. This

transformer based approach can be highly effective when the model is provided with papers containing scientific information. A list of hyper parameters, their values and associated intuitions are listed in the table below:

Hyperparameter	Value	Explanation
Epochs and batch size	7 and 16	As the model is quite complex and the computational power is limited, we chose a low value for these hyperparameters.
Weight decay	0.05	This value ensures the model penalises itself if it is relying too heavily on any particular feature. As we have only provided the model with two features, this value can be low.
Warmup steps	1000	This value ensures that the model dynamically increases the learning rate based on current performance. It will ensure we gain better results.

Table 3. SciBERT model hyperparameter settings

CNN

A CNN is a Convolutional Neural Network which is a type of deep learning model [11]. To train a CNN it requires a dataset split into 3 parts. The train data set, usually the largest part and the validation dataset are both used while training, as the model iterates over each epoch learning from the training data while validating the results using the validation set. At the end of training the test set which has not been seen by the model at all is used to confirm its accuracy. We chose to use a CNN as it is a good baseline deep learning model which performed well in other papers.

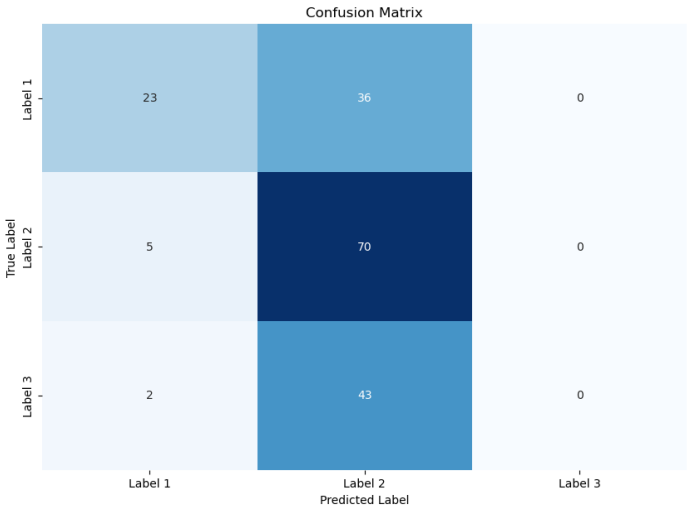
The model learns by randomising the weights of the nodes within the network. It first takes the input as a vector of numbers which is passed through an embedding layer (this layer is skipped when using SCIBERT). Our baseline CNN model has an embedding layer, one convolution layer with 32 filters, one pooling layer and 2 dense layers. Our SCIBERT model is similar, though doesn't include the embedding layer and contains one extra convolution layer also with 32 filters. A CNN will contain Convolution layers to help process the initial data and extract its features, this is done through activation functions whose weights can be altered to affect when a node is fired. The data is then fed into a fully connected layer (dense layer), which contains a large number of nodes which also use activation functions. Finally the information is passed into the output layer (also a dense layer), containing the same amount of nodes as classes, in this case 3. The output layer then decides the most probable class and returns it.

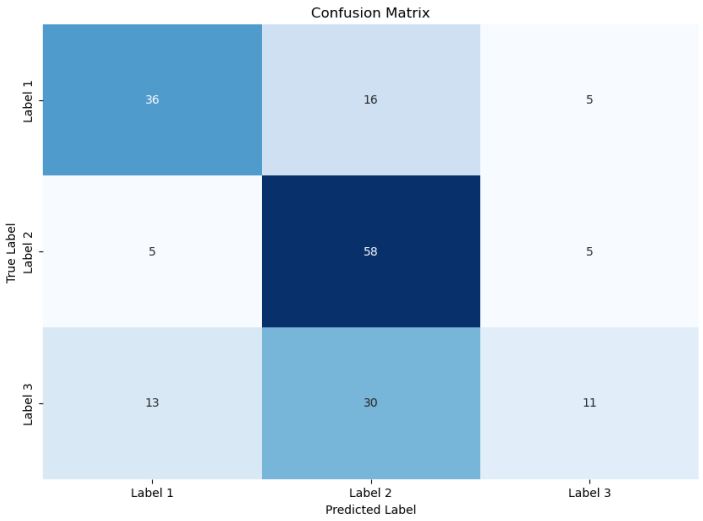
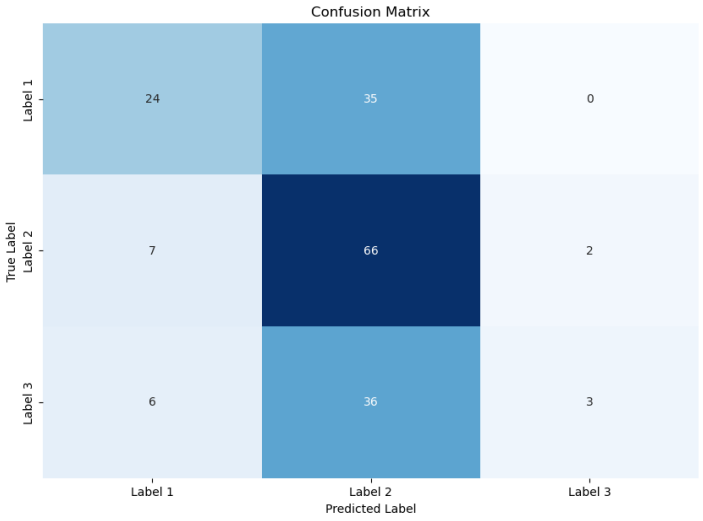
Experiment and Results

Machine Learning Models Results:

Machine Learning Models	Train Accuracy (2.d.p)	Validation Accuracy (2.d.p)	Test Accuracy (2.d.p)	Precision (2.d.p)	Recall (2.d.p)	F1-Score (2.d.p)
SVM (Support Vector Machine)	0.47	0.46	0.52	0.41	0.44	0.38
SVM with SciBERT pre-processing	0.74	0.64	0.59	0.58	0.56	0.54
Random Forest	0.53	0.47	0.52	0.58	0.45	0.41
Random Forest with SciBERT pre-processing	0.75	0.57	0.56	0.57	0.49	0.44

Table 4. Machine Learning Models Results

Machine Learning Models	Confusion Matrix	Results
SVM (Support Vector Machine)		**Label 1 (societal progress):** TP: 23 TN: 113 FN: 36 FP: 7 **Label 2 (scientific progress):** TP: 70 TN: 65 FN: 5 FP: 79 **Label 3 (both):** TP: 0 TN: 134 FN: 45 FP: 0

SVM with SciBERT Pre-Processing		**Label 1 (societal progress):** TP: 36 TN: 104 FN: 21 FP: 18 **Label 2 (scientific progress):** TP: 58 TN: 65 FN: 10 FP: 46 **Label 3 (both):** TP: 11 TN: 115 FN: 43 FP: 10
Random Forest		**Label 1 (societal progress):** TP: 24 TN: 107 FN: 35 FP: 13 **Label 2 (scientific progress):** TP: 66 TN: 33 FN: 9 FP: 71 **Label 3 (both):** TP: 3 TN: 132 FN: 42 FP: 2

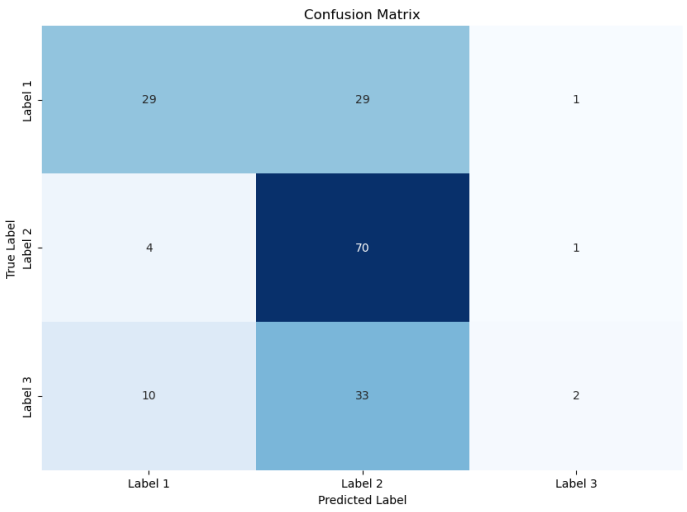
Random Forest with SciBERT Pre-Processing		**Label 1 (societal progress):** TP: 29 TN: 106 FN: 30 FP: 14 **Label 2 (scientific progress):** TP: 70 TN: 42 FN: 5 FP: 62 **Label 3 (both):** TP: 2 TN: 132 FN: 43 FP: 2

Table 5. Machine Learning Models Confusion Matrix

Based on the above results, the SVM model showed relatively low performance across all metrics. The accuracy on the training set was slightly higher than the validation accuracy, indicating some overfitting, but both values were low at 47% and 46% respectively. The test accuracy was slightly higher than the validation and train accuracies at 52%, but was still low, indicating a poor performing model. The precision value was also low at 41%, and the recall was only 3% higher suggesting that the model struggled to correctly classify true positives. This implies that the model had a bias towards predicting negative instances, possibly due to an imbalance in the classes. The f1-score was also low at 38% further indicating that the model to correctly classify the papers. TThe above results satisfy our first research objective as we successfully developed and produced results using the machine learning model SVM to classify the dataset.

Table 1 and 2 demonstrate that whilst the SVM model with SciBERT pre-processing still performed relatively poorly, it was an improvement from the base SVM model. The train, validation and test accuracies were 27%, 18% and 7% higher, signifying significant improvements. The precision, recall and f1-score were all roughly 10% higher, at 58%, 56% and 54% respectively. However, these results are still low indicating that the models performed poorly at correctly classifying the articles, and particularly faced difficulties correctly classifying true positives. The confusion matrix illustrated that the model significantly struggled to correctly classify label 3 (both societal and scientific progress) with only 11 true positive classifications. The further suggested that there was an issue with unbalanced classes. The above results satisfy the project's first questions objective as the SVM model was successfully developed with SciBERT used as a pre-processing method. These results were then compared to the baseline SVM model, discussing the inclusion of

SciBERT which improved the model's performance. This thereby answers the project's first question, indicating that the addition of word embedding models such as SciBERT improved the model's accuracy when being used as a part of the pre-processing.

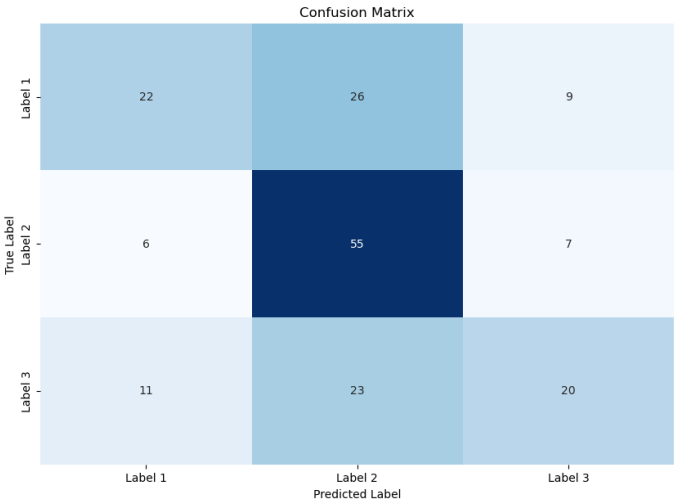
The base random forest model produced similar results to the base SVM model with accuracy values ranging from 47%-52%. The precision value was slightly higher at 58%, indicating that the random forest model was better at correctly classifying true positives. The recall and f1-score values were similar to the base SVM model, being 45% and 41% respectively, demonstrating the model struggled to correctly classify the articles. The confusion matrix demonstrated that the random forest model also struggled to correctly classify the third label, with only 3 true positive values, reinforcing the notion that there was an issue with unbalanced classes. The idea that scientific progress (label 2) formed the dominant class was also evident in the base random forest models results as the true positive value for label two was 42 higher than label 1 and 63 higher than label 3. The above results satisfy our first research objective for the second question as we successfully developed and produced results using the random forest machine learning model to classify the dataset.

This was evident in the confusion matrix as all three labels had more true negative classifications than true positive. This was particularly apparent for societal progress (label 1) and both societal and scientific progress (label 3), with 77 and 130 value differences respectfully. The confusion matrix also illustrated that the model significantly struggled to correctly classify label 3 (both societal and scientific progress) with only 2 true positive classifications. The further suggested that there was an issue with unbalanced classes, with label 2 forming the dominant class. The above results satisfy the project's first questions objective as the random forest model using SciBERT as pre-processing was successfully developed and produced results. These results were then compared to the baseline random forest model, discussing how the inclusion of SciBERT improved the model's accuracy. This assessment answers the project's first question, as it demonstrates that the addition of word embedding models such as SciBERT improved the model's accuracy when being used as a part of the pre-processing.

Deep Learning Models Results:

Deep Learning Models	Train Accuracy (2.d.p)	Validation Accuracy (2.d.p)	Test Accuracy (2.d.p)	Precision (2.d.p)	Recall (2.d.p)	F1-Score (2.d.p)
CNN (Convolutional Neural Network)	1.00	0.61	0.54	0.55	0.52	0.52
CNN with SciBERT Pre-Processing	0.69	0.56	0.53	0.53	0.57	0.55

Table 6. Deep Learning Models Results

Deep Learning Models	Confusion Matrix	Results
CNN (Convolutional Neural Network)		**Label 1 (societal progress):** TP: 22 TN: 105 FN: 35 FP: 17 **Label 2 (scientific progress):** TP: 55 TN: 62 FN: 13 FP: 49 **Label 3 (both):** TP: 20 TN: 109 FN: 34 FP: 16

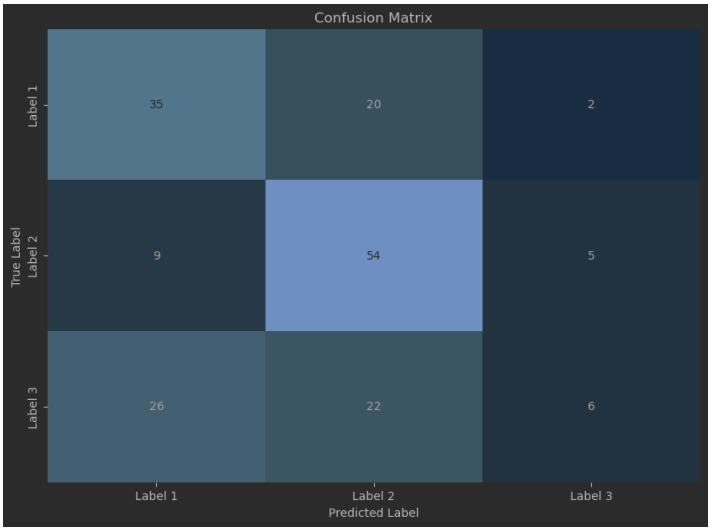
CNN with SciBERT Pre-Processing		**Label 1 (societal progress):** TP: 35 TN: 87 FN: 22 FP: 35 **Label 2 (scientific progress):** TP: 54 TN: 69 FN: 1 FP: 42 **Label 3 (both):** TP: 6 TN: 118 FN: 48 FP: 7

Table 7. Deep Learning Models Confusion Matrix

Looking at tables 3 and 4 the base CNN model produced superior results with a train accuracy of 100% and validation and test accuracies of 61% and 54% respectively. Although the overall accuracy is significantly higher than the SVM and random forest models, this CNN model is grossly overfitting to the input data. The recall and f1-score values of the CNN model were higher than those of the SVM and random forest models, suggesting better performance in classifying articles. However, the CNN model struggled with classifying the first and third labels, implying unbalanced classes. The above results satisfy the project's second questions' second objective as the CNN base model was successfully developed and produced results when classifying the dataset. The project's second question was also answered, as it was concluded that the CNN model produced a higher test accuracy than the base machine learning models.

The study also explored using Sci-BERT as pre-processing for the base CNN model. Surprisingly, this led to worse accuracy values with significantly lower train, validation and test accuracies compared to the base CNN model. Even with Sci-BERT used as pre-processing, the model continued to struggle with the third label, indicating an issue with unbalanced classes. This assessment answers our project's first question, as it demonstrates that the addition of word embedding models such as SciBERT for deep learning models doesn't improve the model's overall accuracy when being used as a part of the pre-processing.

CNN Model Training/Validation vs Epoch Graphs:

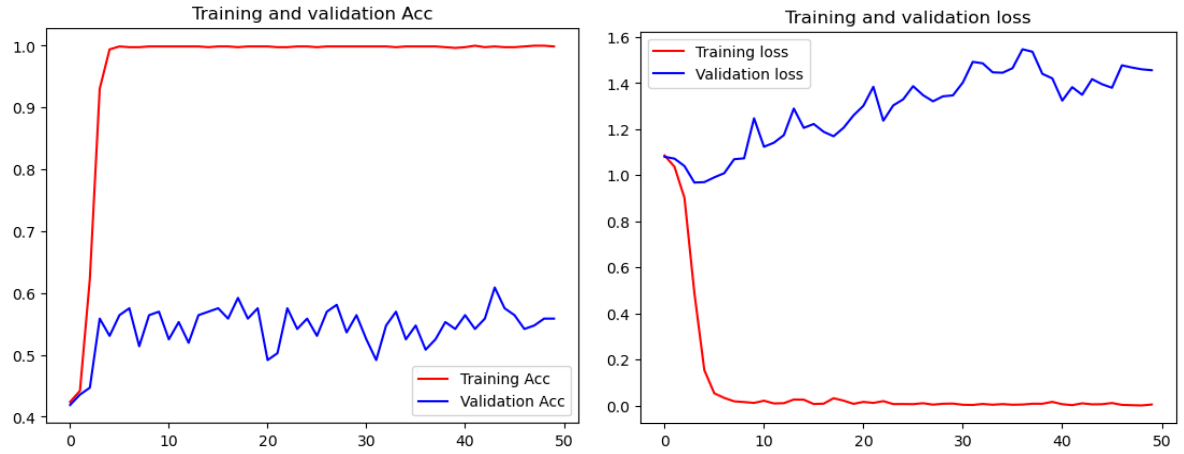


Figure 2. CNN Model Training/Validation vs Epoch

Figure 2 demonstrates the base CNN model was significantly overfitting as after roughly the third epoch the validation accuracy stabilised but the training accuracy continued to increase, reaching a perfect value. The loss graph also illustrates this as the training loss decreases rapidly and reaches close to zero after the fifth epoch but the validation loss slightly increases but does not vary much throughout the fifty epochs.

CNN with SciBERT Pre-Processing Model Training/Validation vs Epoch Graphs:

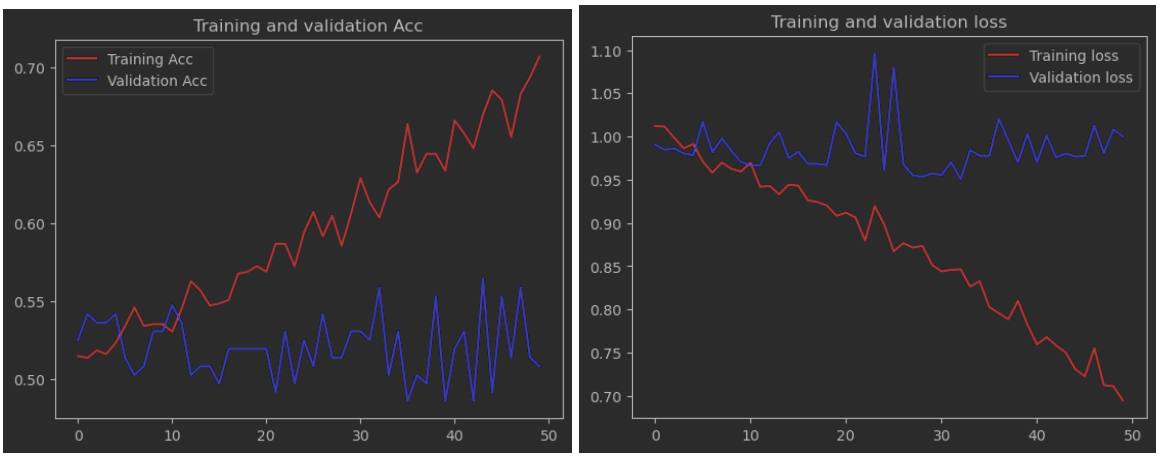


Figure 3. CNN with SciBERT Pre-Processing Model Training/Validation vs Epoch

Figure 3 demonstrates the CNN model with SciBERT applied as pre-processing was significantly overfitting after roughly the tenth epoch as the validation accuracy relatively stabilised but the training accuracy continued to increase. The loss graph also illustrates this as again after around the tenth epoch the training loss continues to decrease but the validation loss begins to even out.

SciBERT Model Results:

	Train Accuracy (2.d.p)	Validation Accuracy (2.d.p)	Test Accuracy (2.d.p)	Precision (2.d.p)	Recall (2.d.p)	F1-Score (2.d.p)
SciBERT	0.97	0.60	0.66	0.68	0.66	0.66

Table 8. SciBERT Model Results

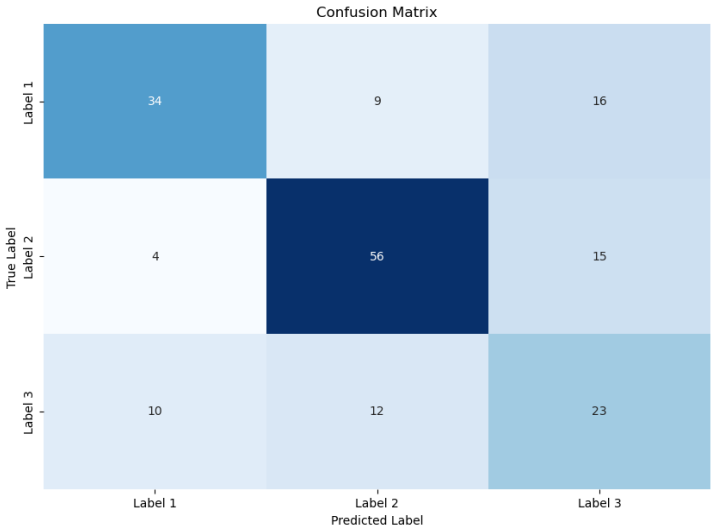
	Confusion Matrix	Results
SciBERT		**Label 1 (societal progress):** TP: 34 TN: 106 FN: 25 FP: 14 **Label 2 (scientific progress):** TP: 56 TN: 83 FN: 19 FP: 21 **Label 3 (both):** TP: 23 TN: 103 FN: 22 FP: 31

Table 9. SciBERT Model Confusion Matrix

Looking at table 5 and 6, the SciBERT model significantly outperformed the base SVM and random forest models in terms of train, validation, and test accuracy, with an accuracy of 97%, 60% and 66% respectively. Although these results are much better than the other models, the model is drastically overfitting to the input data. The precision, recall, and f1-scores of the SciBERT model were also higher, indicating better classification performance. Despite these improvements, the confusion matrix showed that the SciBERT model still struggled with classifying the first (societal progress) and the third label (scientific and societal progress), with the second label (scientific progress) being the dominant class. However, these differences were less drastic than with the previous models.

In conclusion, the SciBERT model was successfully developed and demonstrated superior performance compared to the baseline machine learning and deep learning models in terms of

accuracy, precision, recall, and f1-score values. This effectively answered the project's first question, affirming the utility of the SciBERT model as a classifier.

Result Validation

In this task, we have successfully created a deep learning based classifier, however we must still assess how valid our classifiers are. Artificial intelligence is a highly reliable tool, when we provide it with quality data and perform correct experiments, hence we will be assessing our research validity by first assessing the data quality and then assessing the correctness of each experiment performed.

Dataset

After thorough examination, the provided dataset is most definitely of high quality (See table below). As the task is well defined, and the features are relevant to the task, there is no doubt that this dataset is most definitely valid. However, the main limitation of the dataset is the lack of data. Most deep learning and machine learning based algorithms require a lot of data to generalise, and as we only have 1193 papers in this dataset, it is difficult to create an accurate classifier. Regardless of the size, we believe this provided data is correct and hence valid for this task.

Paper ID	Paper Title	Provided Classification	Verified Classification
16	Barriers to nutritional care for undernourished hospitalised older people	Scientific	Scientific. This is valid
6105	Dispersion of salmon lice in the Hardangerfjord	Scientific	Scientific. This is valid
17099	Mental health problems in adolescents with delayed sleep phase: results from a large population-based study in Norway	Both	Both. This is valid
18069	Resurveying historical vegetation data-opportunities and challenges	Societal	Societal. This is valid
20316	The importance of marine vs. human-induced subsidies in the maintenance of an expanding mesocarnivore in the arctic tundra	Both	Both. This is valid.

Choice of models

As we have chosen two popular machine learning algorithms (SVM and Random Forest) and two relevant deep learning algorithms (SciBERT and CNN) the choice of models are very much valid. The machine learning algorithms provide a good baseline for comparison with the deep learning algorithms, essentially allowing us to confirm which model is ideal for this task. However, a major limitation arises from the fact that there are numerous other models that exist in the machine learning and deep learning space. Some of these models may perform better, however we cannot confirm this without more time and large-scale experiments on a bigger dataset. Hence, for the small scale of this project, we can state that the choice of models are valid.

Parameter settings and adaptability

As explained in the research design section, the parameter choices for each model had a distinct intuition associated with it. These values were determined after multiple experiments, however, if we perform these experiments again, on a larger dataset, we must adjust these parameters as they would not produce ideal results. We must also note that the SciBERT classifier is particularly trained on scientific text, hence we can expect it to classify papers with scientific impact much better than it can classify papers with social impact. This fact is also applied to all models trained using the SciBERT tokenizer as they may also be biased towards the papers with scientific impact. In contrast, the models trained without SciBERT are not necessarily any better as their accuracy on the test dataset is considerably lower in every model except for the CNN model. Hence, we can conclude that if we were to deploy any of these models on a larger scale, we cannot expect them to perform as well due to this crucial fact.

Repeatability verification

When the SVM and Random Forest models are trained using our unique settings, we have ensured that results can be replicated through the use of random state. This parameter allows us to split the dataset in the same way every time, and also output the same results from the model everytime. Contrastingly, the CNN and SciBERT models may produce different results as their innate structure is highly complex and the application of a random seed does not necessarily change this. However, these models can produce somewhat similar results if the dataset is split in the same way every time. As we have ensured this fact, the repeatability of our experiments are most definitely valid.

Discussion and Conclusions

Limitations

Currently, our classification capabilities have certain limitations in terms of accurately categorising scientific papers, indicating the need for further improvements. During the training of our SciBERT model, we encountered several challenges that hindered its performance, with one significant limitation being the limited labelled data.

The lack of substantial amount of labelled data posed a notable constraint in achieving a higher accuracy. Due to this limitation, we were forced to use a smaller sample size of 1193 papers, consisting of both scientific and societal subjects. With limited labelled examples available for training, our model had fewer opportunities to learn and generalise patterns effectively. Consequently, this scarcity of labelled data constrained our model's ability to make more accurate distinctions and classifications of scientific papers. As a result, the achieved accuracy of 0.66 which is relatively low, primarily due to the limited availability of labelled examples. Overcoming this limitation and acquiring a more extensive labelled dataset would be crucial for improving the classification capabilities of our solutions.

On the other hand, for the CNN deep learning model, the effects of class imbalance were somewhat mitigated compared to the machine learning models. This may be due to deep learning models, especially those with convolutional layers like the CNN model used, are generally more robust to class imbalance. The models learn hierarchical and non-linear representations of the data, helping in capturing patterns even in the presence of imbalanced classes. However, the class imbalance still had an impact on the performance of the CNN model and resulted in a lower accuracy for the minority classes and an overemphasis on the majority class. This again could be observed in the confusion matrices, where the true positive values for the minority classes were lower compared to the majority class, particularly for the third label, albeit not as drastically as in the machine learning models.

Another limitation was due to the unfamiliarity with NLP techniques and models, attention was directed towards the building and hyper parameters of the models. We were also unfamiliar with the pre-processing methods needed and focused on only including the necessary features for determining the impact of the research paper. These were determined to be the paper's title and abstract, therefore we did not also include the author's gender. This would have been an important addition as within the given dataset there was some bias with female authors being more likely to be classified as "societal progress".

Future Directions

Addressing class imbalance would be a crucial step to improve the performance of these models for the future. One technique that could be used in the future to address the class imbalance is oversampling, where the minority class samples are replicated or synthesised to match the number of samples in the majority class. This would help provide the models with more examples of the minority class, allowing them to learn more effectively and make better predictions. Oversampling techniques such as Random Oversampling or SMOTE (Synthetic Minority Over-sampling Technique) could be employed to generate synthetic samples based on the characteristics of the minority class, thereby balancing the class distribution.

Another possible solution could instead be undersampling the majority class by reducing the number of samples from the majority class to match the number of samples in the minority class. This approach allows the models to focus more on the minority class during training, preventing the dominance of the majority class and enabling better learning of the patterns within the minority classes. Undersampling techniques like Random Undersampling or Cluster Centroids Undersampling could be used to select a subset of the majority class samples randomly or strategically for training.

K-fold cross validation is a sampling procedure which is generally used to evaluate machine learning models on limited data. The idea of k-folds is to stop two problems which occur when training and testing on a smaller sample of data. The statistical significance which refers to the smaller the test set, means the greater uncertainty of the model. Additionally models with a smaller training set have difficulties learning the pattern of data which can lead to poor predictions. K-fold cross validation addresses the problem by using each data point k times, once as part of the test set and k-1 as part of the training set. This ensures that the model maximises the learning potential of the data.

Another approach would be to use a combination of the oversampling and undersampling techniques (hybrid sampling). This approach would aim to balance the class distribution by oversampling the minority class and undersampling the majority class simultaneously. By doing so, it would provide the models with a balanced training dataset that could better capture the characteristics of all classes.

By employing these techniques in the future, the models could better capture the underlying patterns and characteristics of all classes, leading to more accurate predictions and improved performance on the minority classes.

Conclusion/ impact of study:

We propose an AI-based classification system to manage the greatly increasing volume of scientific papers, focusing on the contributions to scientific, societal progress and overlap between the research. We used 4 models for this task: SVM, Random Forest, SciBERT, and CNN. The dataset contained 1193 research papers with 15 features each, where we used the title and abstract of each paper as they were the most informative features for determining its impact. Throughout this study, we encountered issues with the limited data, also due to the collective inexperience in the NLP field, alot of concepts were new which halted initial design choices and planning of the project. From our results, we unfortunately did not reach a satisfactory conclusion, we suggest future research directions which theoretically should improve future iterations of this project.

References

[1] Fire, M. (2019) *Over-optimization of academic publishing metrics: Observing Goodhart's ...*, *Over-optimization of academic publishing metrics: Observing Goodhart's Law in action*. Available at: https://www.researchgate.net/publication/333487946_Over-optimization_of_academic_publishing_metrics_Observing_Goodhart%27s_Law_in_action (Accessed: 29 May 2023).

[2] Bichitrananda, B., Kumaravelan, G. and Kumar, P. (2023) *Performance Evaluation of Deep Learning Algorithms in Biomedical Document Classification*, *IEEE Explore*. Available at: https://ieeexplore.ieee.org/abstract/document/9087283/ (Accessed: 01 June 2023).

[3] Hande, A. *et al.* (2021) *Domain identification of scientific articles using transfer learning and ensembles*, *SpringerLink*. Available at: https://link.springer.com/chapter/10.1007/978-3-030-75015-2_9 (Accessed: 24 March 2023).

[4] Huang, C. (2023). "Review of the project proposal - Classification of research papers". UTS

[5] Du, J. *et al.* (2019) *ML-Net: multi-label classification of biomedical texts with deep neural networks*, *Academic.oup.com*. Available at: https://academic.oup.com/jamia/article/26/11/1279/5522430?login=true (Accessed: 24 March 2023).

[6] Allenai (2020) *Allenai/scibert: A bert model for scientific text.*, *GitHub*. Available at: https://github.com/allenai/scibert (Accessed: 13 March 2023).

[7] Cheng, R. (2020) *NLP text preprocessing with NLTK | Towards Data Science*, *Text Preprocessing With NLTK*. Available at: https://towardsdatascience.com/nlp-preprocessing-with-nltk-3c04ee00edc0 (Accessed: 28 May 2023).

[8] Gandhi, R. (2018) Support Vector Machine — introduction to machine learning algorithms ..., Support Vector Machine — Introduction to Machine Learning Algorithms. Available at: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 (Accessed: 28 May 2023).

[9] Yiu, T. (2019) *Understanding random forest - towards data science*, *Understanding Random Forest*. Available at: https://towardsdatascience.com/understanding-random-forest-58381e0602d2 (Accessed: 28 May 2023).

[10] Gupta, Y. (2021) *Text classification with SciBERT*, *Text Classification with SciBERT*. Available at: https://guptayash2010.medium.com/text-classification-with-scibert-a285d2f2db06 (Accessed: 28 May 2023).

[11] Choubey, V. (2020) *Text classification using CNN*. Available at: https://medium.com/voice-tech-podcast/text-classification-using-cnn-9ade8155dfb9 (Accessed: 28 May 2023).