Fair Model for Disinformation Detection with respect to global North and global South

Midterm Project Evaluation Report

by

Sujit Mandava
(112001043)

COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY PALAKKAD

CERTIFICATE

*This is to certify that the work contained in the project entitled "**Fair Model for Disinformation Detection with respect to global North and global South**" is a bonafide work of **Sujit Mandava** (**Roll No. 112001043**), carried out in the Department of Computer Science and Engineering, Indian Institute of Technology Palakkad under my guidance and that it has not been submitted elsewhere for a degree.*

Dr.Sahely Bhadra

Assistant/Associate Professor

Department of Computer Science & Engineering

Indian Institute of Technology Palakkad

Contents

1 Introduction **1**

1.1 Progress in Term-1 1

1.2 Problem Statement 2

1.3 Organization of The Report 2

2 Work Done **3**

2.1 Current Dataset Division 3

2.2 Tests 4

2.3 Models Used 4

2.3.1 Model A 4

2.3.2 Model B 5

2.3.3 Model C: FNDNet[1] 5

2.4 Test Set Up 6

2.4.1 Test 1 7

2.4.2 Test 2 7

2.4.3 Test 3 7

2.5 Test Results and Observations 8

2.5.1 Test 1 8

2.5.2 Test 2 8

2.5.3 Test 3 9

3 Further Work and Conclusion **15**

References **17**

Chapter 1

Introduction

1.1 Progress in Term-1

The problem statement of the project can broadly be divided into two parts: confirming the hypothesis that popular disinformation detection techniques show a bias towards the Global North and creating a disinformation detection system that overcomes the bias if any.

In the first half of the project, we tested a few popular disinformation detection methods to see if any bias towards the global north could be seen. After testing the models we concluded that there was indeed some amount of bias we observed, but upon further examination of the test results, we determined that the evidence was not sufficient enough to confirm the presence of said bias.

We also concluded that there was comparatively a lesser amount of data originating from the Global South compared to that from the Global North, which could be a possible cause for the hypothetical bias towards the Global North.

Another possible area where a bias can arise is the design of the neural network and detection algorithm itself. Since most disinformation detection research is currently focused on the Global North, certain elements in the algorithm itself may introduce a bias toward data originating from the Global North.

1.2 Problem Statement

Our current goal can be divided into two steps. The first step is to establish that current disinformation detection techniques show a bias towards data originating from the Global North and to identify the source of the bias if it exists. The next step is to create a disinformation detection system that performs equally well on data from both the Global North and the Global South and is at par with current state-of-the-art disinformation detection techniques.

1.3 Organization of The Report

Chapter 1 gives us a brief overview of the work done in the project in the previous term and the following report's overview. Chapter 2 gives us an insight into the work done in this semester. It goes over the tests done, the motivation behind them, and the observations from the same. Chapter 3 talks about the future work and serves as a conclusion for the report,

Chapter 2

Work Done

2.1 Current Dataset Division

We have selected the following datasets to proceed with further tests to establish the existence of said bias. We can see the discrepancy in the amount of news articles we have

Dataset Names	Global North		Global South	
	Real	Fake	Real	Fake
ISOT Fake News[2][3]	16544	23481	4873	-
FakeNewsIndia[4]	N/A	N/A	0	4843
IND[5]	N/A	N/A	200	0
ToI	N/A	N/A	668	0
Total	16544	23841	5741	4843

Table 2.1 Dataset Breakdown

gathered from the Global North and the Global South. A web crawler was implemented to gather news articles about the Global South, to overcome the large difference in the number of articles. However, the current version of the web crawler is a simple tool that extracts news articles from the Times of India website. It can currently extract about 85 articles daily and must be run manually daily.

2.2 Tests

The following tests were deemed necessary to determine if the hypothesis is true:

1. Training the selected models on Global North data and testing it on both Global North and Global South Data
2. Training the selected models on Global South data and testing it on both Global North and Global South Data
3. Training the selected models on an equal distribution of Global North and Global South data and testing it on both Global North and Global South Data

Test 1 emulates how current disinformation detection systems would work, as most systems are trained almost exclusively on Global North data. Test 2 aims to confirm the hypothesis that a lack of data originating from the Global South plays a major role in the difference in performance. Test 3 aims to check if any bias originates from the models' algorithm and neural network design. A balanced training dataset means that any performance difference that may arise is solely due to the algorithm design.

2.3 Models Used

2.3.1 Model A

```
|-----------------------------------------------------|
| Layer                   | Output Shape     | Param #    |
|-------------------------|------------------|------------|
| embedding (Embedding)   | (None, 1000, 100)| 3,000,000  |
| conv1d (Conv1D)         | (None, 997, 128) | 51,328     |
| max_pooling1d           | (None, 249, 128) | 0          |
| dropout                 | (None, 249, 128) | 0          |
```

flatten	(None, 31872)	0
dense	(None, 128)	4,079,744
dropout	(None, 128)	0
dense	(None, 1)	129

2.3.2 Model B

Layer	Output Shape	Param #
embedding (Embedding)	(None, 1000, 100)	3,000,000
flatten	(None, 100000)	0
dense	(None, 1)	100001

2.3.3 Model C: FNDNet[1]

Layer	Output Shape	Param #
embedding (Embedding)	(None, 1000, 100)	3,000,000
conv1d_1 (Conv1D)	(None, 998, 128)	38,528
conv1d_2 (Conv1D)	(None, 997, 128)	51,328
conv1d_3 (Conv1D)	(None, 996, 128)	64,128
max_pooling1d_1	(None, 199, 128)	0
max_pooling1d_2	(None, 199, 128)	0
max_pooling1d_3	(None, 199, 128)	0
concatenate(Concatenate)	(None, 597, 128)	0

```
| conv1d_4 (Conv1D)       | (None, 593, 128) | 82,048     |
| max_pooling1d_4         | (None, 118, 128) | 0          |
| conv1d_4 (Conv1D)       | (None, 114, 128) | 82,048     |
| max_pooling1d_4         | (None, 3, 128)   | 0          |
| flatten_2 (Flatten)     | (None, 384)      | 0          |
| dense_3                 | (None, 128)      | 49,240     |
| dropout_2               | (None, 128)      | 0          |
| dense_4                 | (None, 1)        | 129        |
|-------------------------------------------------------|
```

2.4 Test Set Up

For all three tests, the following training and testing pipeline is used[1]:

1. The dataset is first preprocessed using the Keras library. It is tokenized and the sequenced tokens are used to generate a GloVe embedding.

2. Dataset is split into training data and test data. The training data is further split into training and validation data.

3. The three models are trained using the GloVe embedding and the training data.

4. The performance is then tested on the unseen test data and the performance metrics are tabulated.

5. Finally, the performance of the model is tested on the entire dataset. The amount of data chosen from the entire dataset is varied from a minimum to the maximum available data. The performance metrics are then plotted against the amount of data in the test dataset.

[1] Notebooks

2.4.1 Test 1

The data used is from the ISOT dataset[2][3], which includes all the Global North news as previously mentioned in the table. The number of real news articles is less than that of fake news, thus we use the SMOTE technique to oversample the real news in the training data, after splitting the dataset into train and test data and converting the training data into tokenized sequences.

2.4.2 Test 2

The data used is gathered from FakeNewsIndia[4], IND[5], ISOT[2][3], and news scrapped from ToI. The number of real news articles is more than that of fake news, thus we use the SMOTE technique to oversample the fake news in the training data, after splitting the dataset into train and test data and converting the training data into tokenized sequences.

2.4.3 Test 3

A combination of the datasets used in test 1 and test 2 is used. There is a large disparity in the amount of data available when the Global North is compared to the Global South (40,385 vs 10584). Thus, we have to combine the undersampling of the Global North data and the oversampling of the Global South data. We oversample the Global South data to approximately 12,000 entries and undersample the Global North data to the same amount. However, this may result in inaccurate findings as a large amount of Global North data remains unseen whereas the Global South data has been oversampled. The models may show trends of overfitting to the Global South data.

2.5 Test Results and Observations

2.5.1 Test 1

The results of test 1 were in line with our predictions. The three models performed exceptionally well with high precision and recall on test data from the Global North, whereas the performance on Global South data was poor.

2.5.2 Test 2

As expected, the three models performed better in classifying Global South data as compared to the Global North data. However, we can make certain observations upon comparing the exact precision and recall values obtained from test 1 and test 2:

1. The precision and recall of the three models when tested against Global North data in test 1 is higher than that of the models when tested against Global South data in test 2.

2. The precision and recall of the three models when tested against Global South data in test 1 is higher than that of the models when tested against Global North data in test 2.

A key difference between the two tests lies in the number of articles present in the training datasets. The size of the training dataset in test 1 is almost double that of test 2. This is reflected in observation 1, where Global North-trained models perform better on Global North data (similar to the training data) compared to the performance of Global South models on Global South data. This difference also plays a role in observation 2; since the sample size for test 1 is much larger than that of test 2, Global North models are more capable of identifying and classifying new, unseen data and thus perform better on Global South data compared to Global South models on Global North data.

2.5.3 Test 3

We observe that the three models appear to perform at a similar level on both the Global North and the Global South data. The slightly better performance in the case of the Global South data (in models A and B) can be explained by our initial prediction, i.e, the oversampling of Global South data and undersampling of Global North data would result in lost Global North context and cause the models to overfit to the Global South Data.

Model	**GN Unseen Test Data**		
	Precision	Recall	F1
Model A	0.949	0.978	0.963
Model B	0.956	0.986	0.971
Model C[1]	0.993	0.995	0.994
Model	**GS Unseen Test Data**		
	Precision	Recall	F1
Model A	0.978	0.945	0.961
Model B	0.991	0.965	0.978
Model C[1]	0.993	0.995	0.994
Model	**Equal Split Unseen Test Data**		
	Precision	Recall	F1
Model A	0.862	0.843	0.852
Model B	0.950	0.934	0.942
Model C[1]	0.976	0.977	0.977

Table 2.2 Evaluation Metrics for Different Models

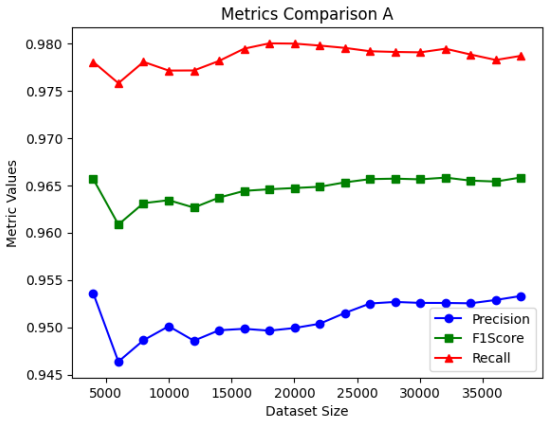
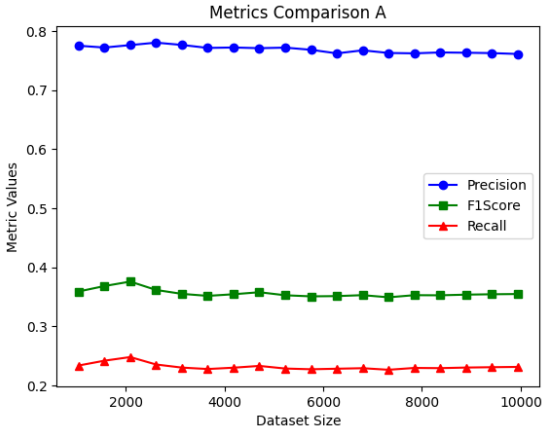
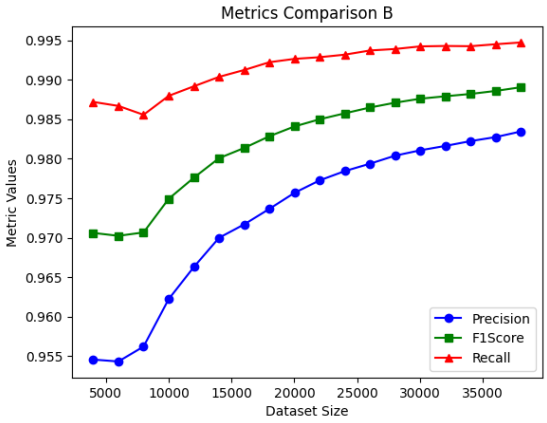
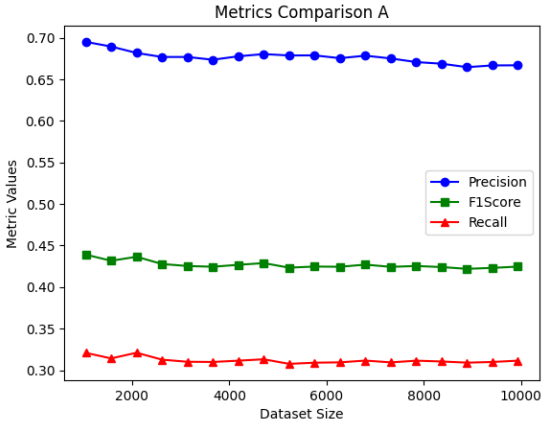
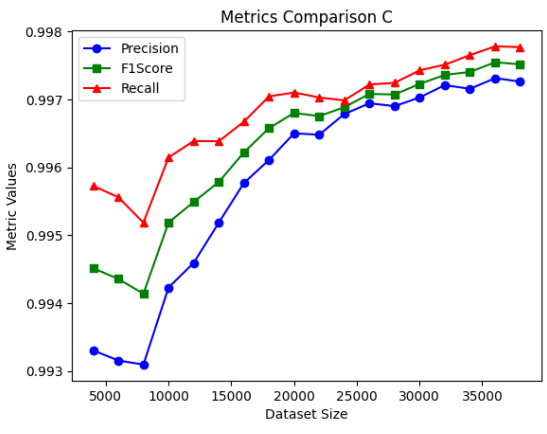
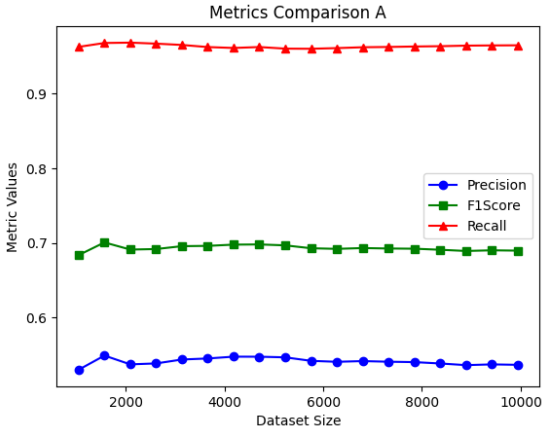
	GN Test Data	GS Test Data
Model A		
Model B		
Model C[1]		

Table 2.3 Comparison of GN Models against GN and GS Test Data

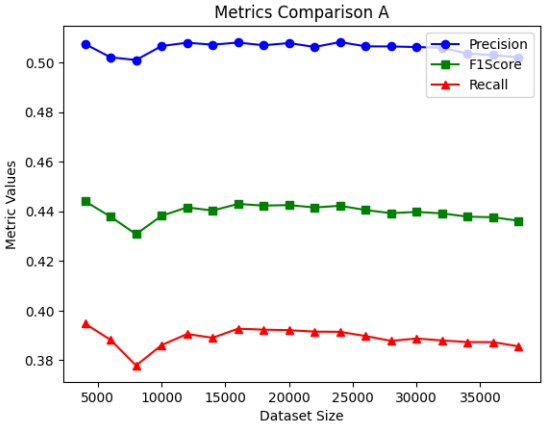
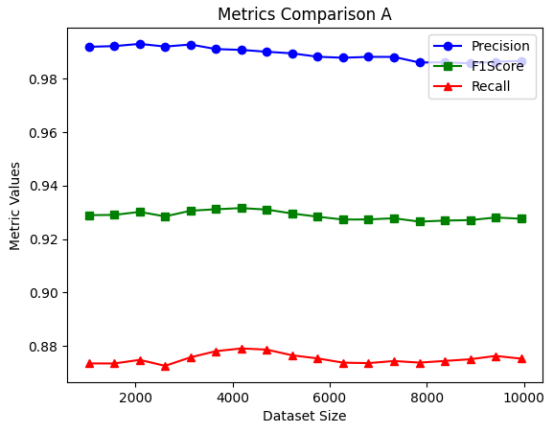
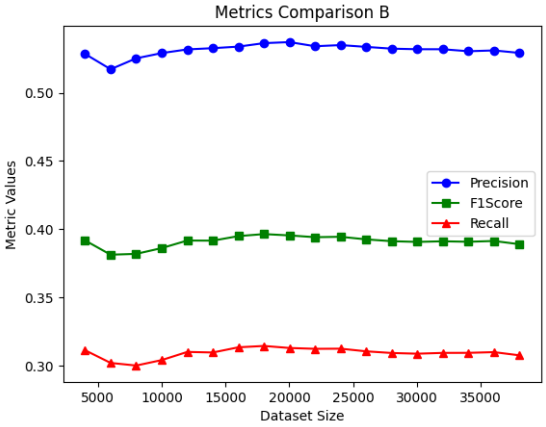
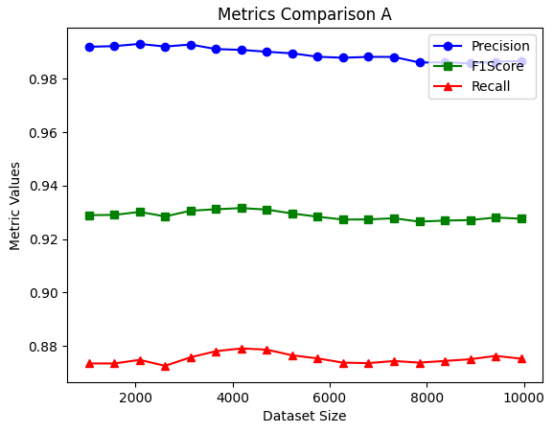
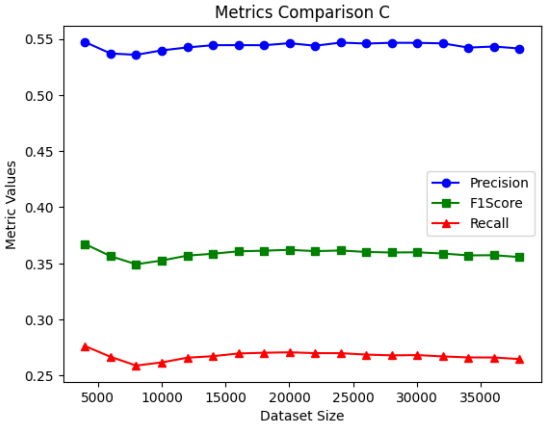
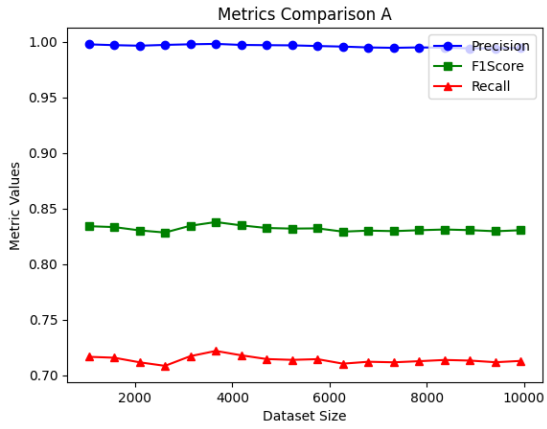
	GN Test Data	GS Test Data
Model A		
Model B		
Model C[1]		

Table 2.4 Comparison of GS Models against GN and GS Test Data

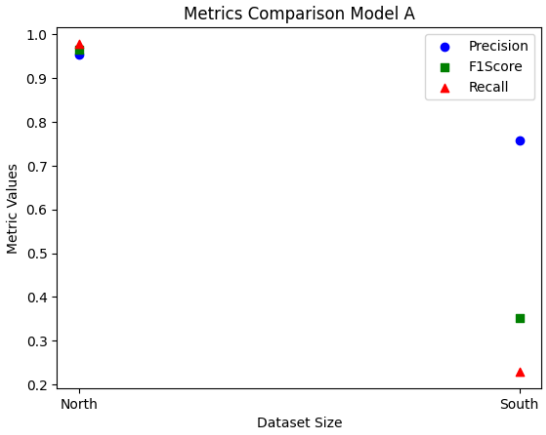
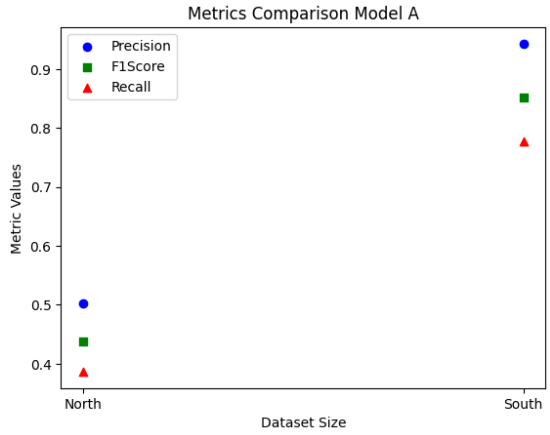
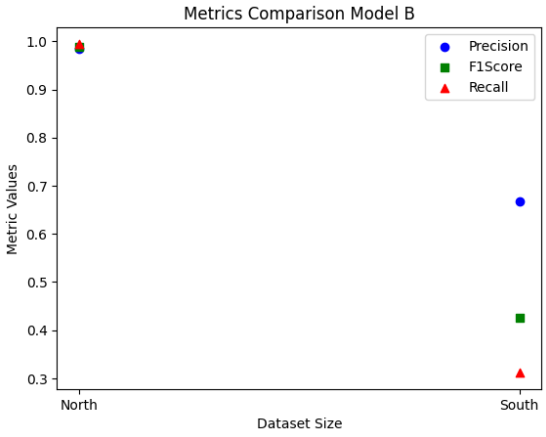
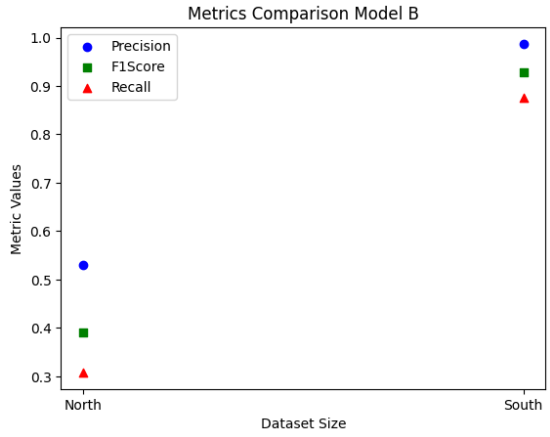
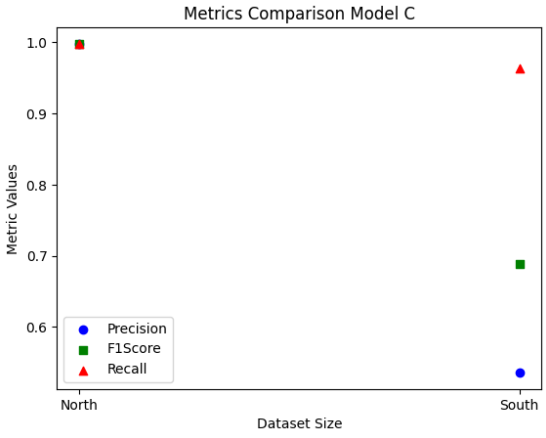
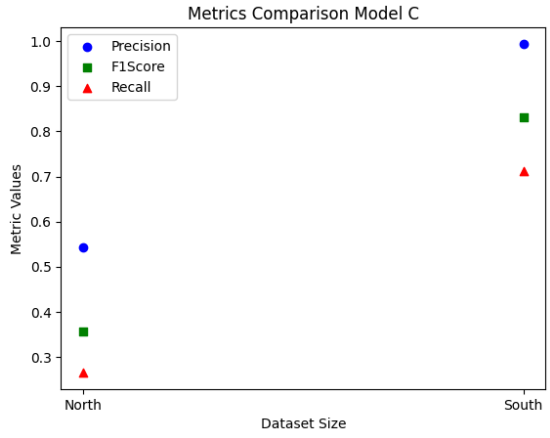
	GN Training Data	GS Training Data
Model A		
Model B		
Model C[1]		

Table 2.5 Performance comparison of models on Global North Data vs Global South Data (Tests 1 and 2)

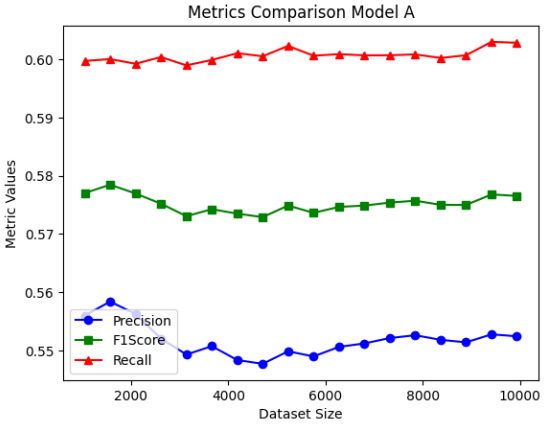
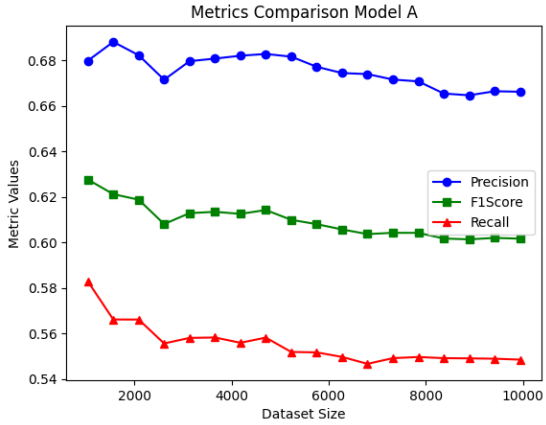
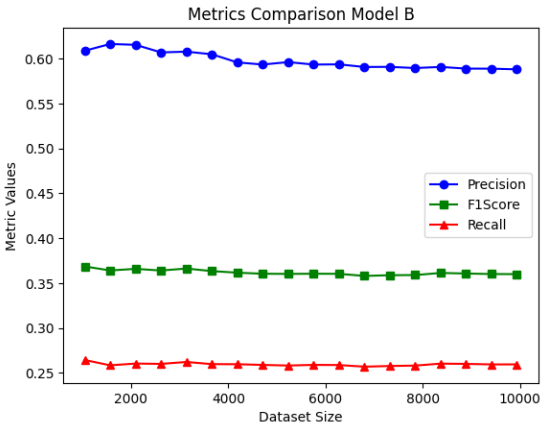
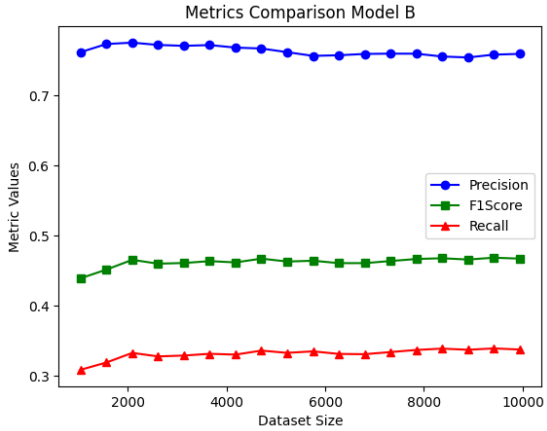
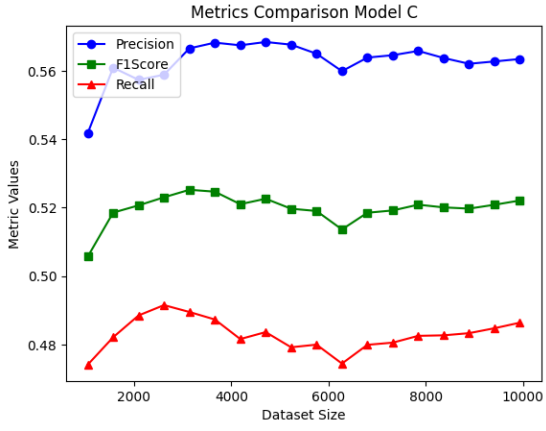
	GN Test Data	GS Test Data
Model A		
Model B		
Model C[1]		

Table 2.6 Comparison of Equal-Data Models against GN and GS Test Data

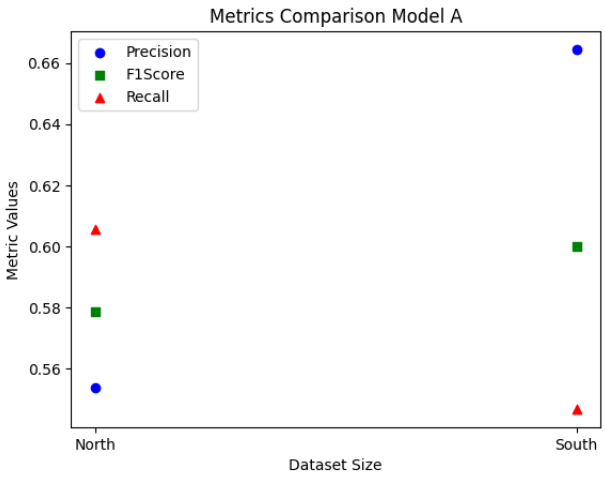
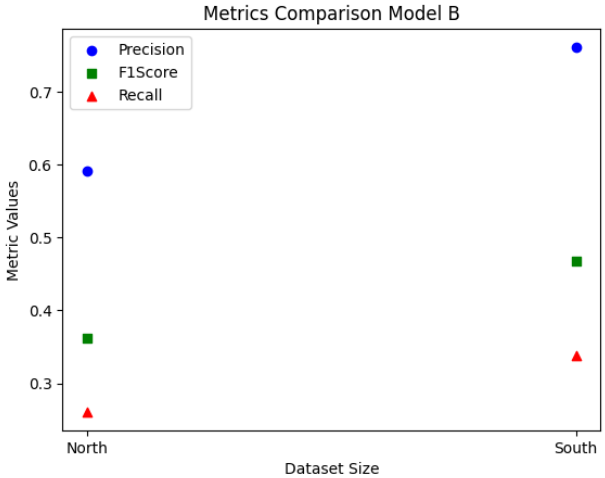
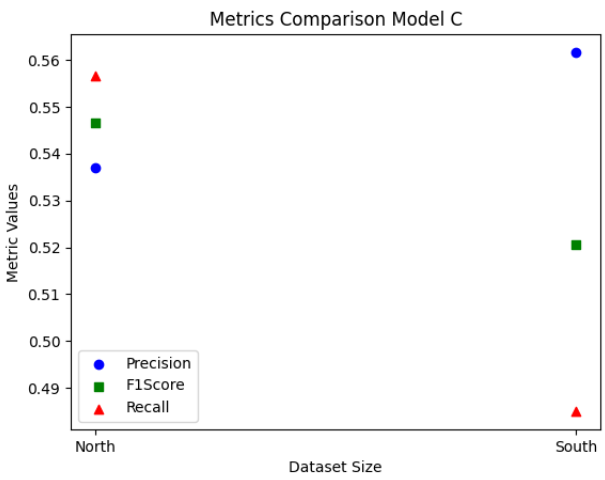
	Equal Training Data
Model A	
Model B	
Model C[1]	

Table 2.7 Performance comparison of models on Global North Data vs Global South Data (Test 3)

Chapter 3

Further Work and Conclusion

It is evident from the results of test 1 that the performance of current disinformation detection models is better on data originating from the Global North. However, it is safe to assume that the bias originates from a smaller dataset originating from the Global South. This can be inferred from the observations made from the results of tests 1 and 2.

We also observe that the models perform at similar capacities when tested against Global North and Global South data in test 3. Thus, we can say that there is no bias being induced due to the architecture of the models or the chosen algorithms. The slightly better performance in the case of Global South data here can be attributed to the models overfitting the Global South data as they have been oversampled.

We can therefore conclude safely that any bias shown by disinformation detection systems towards the Global North, if observed, can be attributed to the lack of sufficient data and context related to the Global South.

The upcoming work for this project will focus on creating a fair disinformation detection system. The first task will mainly aim to isolate individual causes for the bias and aim to tackle them one by one. All the potential sources of the bias must be dealt with to create a fair disinformation detection system. The state-of-art models will then be analyzed to gather inspiration for the new detection system. Multiple directions such as including new

features will be explored to boost the model's performance.

One such idea worth exploring is introducing LLMs into the pipeline. Large Language Models, or LLMs, are transformer models trained on large amounts of text data and capable of learning and generating human language. These pre-trained models that are trained on massive amounts of data found at various sources across the Internet can generate and process natural language and learn a lot of world knowledge through the training process. LLMs can also be fine-tuned as needed to boost their performance. They can thus be leveraged to add more helpful context to the disinformation detection system such as the source's trustability, the text's tone, area of origin, etc.[6]

Multiple studies talk about the difference between real news and fake news. One such research paper talks about the difference in writing styles and linguistic characteristics between real and fake news. Fake news tends to be more polarizing and subjective, whereas real news is much more objective and typically is accompanied by verified facts and statistics. This information can be extracted by LLMs from texts owing to the extraordinary capability of processing natural languages.

Future work will aim to assess the functioning of LLMs in the mentioned direction. Pairing current state-of-the-art techniques with LLMs providing additional context can boost the performance of these disinformation detection pipelines. Since LLMs are trained on a large amount of data, we can assume that they have equal and sufficient context regarding both the Global North and the Global South, thus tackling a major problem that has been identified from the very beginning: insufficient data related to the Global South.

References

[1] R. K. Kaliyar, A. Goswami, P. Narang, and S. Sinha, "Fndnet – a deep convolutional neural network for fake news detection," *Cognitive Systems Research*, vol. 61, pp. 32–44, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389041720300085

[2] H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using n-gram analysis and machine learning techniques," in *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1*. Springer, 2017, pp. 127–138.

[3] ——, "Detecting opinion spams and fake news using text classification," *Security and Privacy*, vol. 1, no. 1, p. e9, 2018.

[4] A. Dhawan, M. Bhalla, D. Arora, R. Kaushal, and P. Kumaraguru, "Fakenewsindia: A benchmark dataset of fake news incidents in india, collection methodology and impact assessment in social media," *Computer Communications*, vol. 185, pp. 130–141, 2022.

[5] R. Suharshala, A. Kadan, M. P.Gangan, and L. V L, "Online news popularity prediction before publication: effect of readability, emotion, psycholinguistics features," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, pp. 539–545, 06 2022.

[6] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.