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

Chapter 1

Introduction

1.1 Background

In today's interconnected world, information serves as the lifeblood of societies, influencing our choices, beliefs, and actions daily. It guides our understanding of the world and empowers us to make informed decisions in almost every realm of society. However, a shadow looms over this information landscape in the form of disinformation; false or misleading content deliberately spread to deceive, manipulate, or sow discord. The adverse impacts of disinformation on society are profound and far-reaching. It erodes trust in institutions, fuels division among communities, and poses a severe threat to the very foundations of democracy. Recognizing the gravity of this issue, addressing disinformation, and developing effective detection tools have become essential challenges of our times.

It is crucial to acknowledge that the impact and dynamics of disinformation may differ across regions, particularly between the global North and the global South. Existing tools for detecting disinformation, which are often developed in and tailored to Western contexts, may not perform optimally when applied to samples from the global South. Factors such as linguistic diversity, cultural nuances, and variations in media ecosystems can pose unique challenges to disinformation detection in different regions. There is a growing need to develop more inclusive and context-sensitive solutions that address said nuances in the

global South. Recognizing these discrepancies and working toward more unbiased solutions is vital to the fight against disinformation and the preservation of information integrity worldwide.

1.2 Problem Statement

The problem statement can be divided into three broad sections: analysis of current disinformation detection models, dataset creation, and fair model generation.

1.2.1 Analysis of Disinformation Detection Models

The first part is to find and analyze the performance of current state-of-the-art disinformation detection techniques and models on fake news related to both the Global North and Global South. The objective is to test the hypothesis that current disinformation detection techniques perform better on news from the Global North and do not perform as well on news from the Global South, due to most fake news datasets exhibiting a large bias towards the Global North. All the models chosen employ content-based detection techniques. If the hypothesis holds, the task is to try and identify the reasons this bias is shown.

1.2.2 Dataset Creation

The second is to find real and fake news data related to the Global North and South from various sources and form a fair dataset to test the chosen disinformation detection models. One hurdle to overcome is introducing enough data related to the Global South. The topics covered by the articles in the datasets must be similar, to prevent any bias due to the topic of the content.

1.2.3 Fair Model Generation

The final goal is to create a fair disinformation detection system that mitigates the biases observed from our analysis of the earlier models. The model will mainly employ content-

based detection methods, while possibly including elements of knowledge-based methods to boost performance.

1.3 Organization of The Report

Chapter 1 provides background regarding the problem and the motivation for solving it, and gives us an idea of the problem statement. Chapter 2 gives a brief overview of the literature surveyed. It gives us insights into current research related to disinformation detection and a few state-of-the-art techniques that are being studied as a part of this project. Chapter 3 deals with popular datasets that are currently used for disinformation detection models, and identifies potential issues with current models. Chapter 4 deals with the experimental set up for the analysis of the chosen state-of-the-art techniques and the results of the same. Finally, chapter 5 concludes the report and gives an overview of the future work required as a part of this project.

Chapter 2

Literature Review

Disinformation detection has gained a lot of attention in the past few years, owing to the increase in the amount of fake news being circulated and an increase in awareness in the general public. Most research related to this domain can be divided into three broad categories: content-based detection, social network-based detection, and knowledge-based detection.[2] This project delves deeper into the content and knowledge-based detection methods.

2.1 Content-Based

Content-based detection methods aim to extract features from the semantics involved in the text and classify the articles based on the same. Studies have shown that fake news tends to be more provocative and subjective while real news is more specific and objective. The length of the content, the total number of words, abbreviations, and other similar linguistic features play a major role in determining the nature of the content.[2]

2.2 Knowledge-Based

Knowledge-based methods aim to extract the key information and facts from the content and verify them against some existing database. This can be done in two ways: traditional manual verification and automatic verification. Manual verification is highly accurate but suffers from extremely low efficiency. Automatic methods use natural language processing and machine learning techniques to internalize information using knowledge databases and graphs. Fact verification entails comparing the news with said knowledge graph, based on which the nature of the content is determined.[2]

2.3 Social Network Based

Content-based and knowledge-based methods, while popular, can sometimes be misled by replicating the linguistic cues of real news and using vague facts that can pass as real. Auxiliary data, such as social background, propagation in social networks and public discourse provides valuable information that can improve classification accuracy.[2]

2.4 Current Research on Content-Based Detection

Several approaches have been employed to identify and mitigate the impact of disinformation. Machine learning-based methods, including supervised and unsupervised learning, have gained prominence in distinguishing between legitimate and deceptive content. Classification models like Support Vector Machines, Naive Bayes' classifiers, and deep learning models that use RNNs, CNNs, and LSTM networks are prevalent in this domain. Deep learning models, however, exhibit better performance by an average of around 6% over traditional machine learning models.[2][3] Natural language processing (NLP) techniques contribute immensely to analyzing linguistic patterns, sentiment, and semantic structures to identify anomalies indicative of disinformation. Deep learning models often use pre-learned word embeddings as a feature in their decision-making. Fine-tuning pre-trained

models to achieve high accuracy in a specific task is also an idea that has gained traction in recent years due to its immense success in boosting performance[4].

A path-breaking research paper that explored the idea of using deep learning with word embeddings is "Fake news detection: A hybrid CNN-RNN based deep-learning approach" (Jamal et al., 2020).[5] The authors proposed an approach that combines both CNNs and LSTMs to extract local features as well as learn long-term dependencies to classify the text. This approach scored better than state-of-the-art detection methods and has been used as a state-of-the-art comparison for newer disinformation detection models. Some other research papers that explore the idea of combining word embeddings with deep learning models include "FNDNet – A deep convolutional neural network for fake news detection"[6]. "DISCO: Comprehensive and Explainable Disinformation Detection"[1][Appendix A] is another research paper with a slightly different, novel approach. The proposed model combines graph algorithms with word embeddings to extract new features for the neural classifier.

Chapter 3

Datasets

The datasets used by these models can be single-modal or multi-modal datasets. Single-modal datasets utilize data from a single source such as texts, audio, images, etc. Most single-modal datasets involve using text to determine the nature of the content. Multi-modal datasets combine multiple types of data to possibly make a more informed decision regarding the data. While single-modal datasets are much more efficient, they may suffer from a lack of context. However, the performance boost from utilizing multiple types of income may not be worth the operational overhead required.[2]

An overview of popularly cited datasets in disinformation detection research points to a glaring problem: the lack of data related to the Global South compared to the Global North. Therefore to compare the performance of the chosen models on data from the Global North and Global South, the need to introduce more data from the Global South arises. Datasets like FA-KES[7] and FakeNewsIndia[8] are some popular, reputed datasets containing data that pertains to fake news in the Global South. However, the size of these datasets is lacking severely when compared to the previously mentioned datasets.

This lack of data poses a problem to this study, which can be supported with the help of an example. Consider the model proposed by Jamal et al, 2019[5]. This model is trained on a combination of the ISOT[9][10] and FA-KES[5] datasets. There is a disparity in the

amount of data on the Global North and Global South. The performance score on the ISOT dataset[9][10] is nearly 0.99, the same for the FA-KES[5] dataset is close to 0.60[5]. While the performance of the model on the Global South data in the ISOT dataset[9][10] is unknown, the poor performance in the FA-KES[5] dataset supports the idea that biased datasets result in a biased model. This is precisely the problem we aim to solve, and we explore this comparison in the upcoming chapters.

Chapter 4

Experiments and Results

4.1 Dataset

This study combines two fake news datasets: ISOT Fake News Dataset[9][10], FakeNewsIndia: A baseline Dataset of fake news incidents in India[8], and the FA-KES Dataset. The ISOT Fake News dataset[9][10] consists of two CSV files. The first file, True.csv, contains 21,417 real articles gathered from verified and well-reputed news outlets like Reuters. The second file, False.csv, contains 23,481 articles collected from various unreliable news sites flagged by fact-checking organizations like Politifact and Wikipedia. The articles cover a wide range of topics but mostly focus on world news and political news. However, there is a clear disparity in the number of articles about the Global South. In 21,417 true articles, only 4873 contain news related to the Global South, whereas in 23,481 fake articles, less than a thousand are related to the Global South.[9][10]

The FakeNewsIndia Dataset[8] was chosen to overcome this obvious lack of articles related to the Global South. This dataset contains 4843 articles from various news sources in India, belonging to the Global South. Most of these articles are related to politics, which fits well with the ISOT Dataset[9][10]. Another dataset related to the Global South is the FA-KES[5] dataset, which contains articles related to the Syrian war. This dataset contains an additional 807 articles, of which 426 are true, and the remaining 376 are fake.

The final dataset we use for testing looks like the following:

Dataset Names	Global North		Global South	
	Real	Fake	Real	Fake
ISOT Fake News	16544	23481	4873	$\leq 4\%$
FakeNewsIndia	N/A	N/A	0	4843
FA-KES	N/A	N/A	426	376
Total	16544	23841	5299	5219 $\pm$100

Table 4.1 Dataset Breakdown

Every article is annotated with the location of interest, i.e., whether the article talks about the Global North or Global South. For True.csv, FakeNewsIndia[8], and FA-KES[5] this was fairly straightforward as the location related to the topic was already available (For True.csv, the location of interest was attached to the start of the article). For Fake.csv however, the location was extracted by running a simple Python script. This allows us to check whether the performance of the disinformation detection systems differs between the two regions.

4.2 Testing and Results

We use the following models and test their performance on the dataset generated in the previous section:

- DISCO
- Logistic Regression with GloVe embeddings (link)
- FNDNet (link)

Since both model 2 (Logistic Regression with GloVe embeddings) and model 3 (FNDNet) were trained on ISOT, we trained the models using the ISOT dataset using the same architecture and hyperparameters prescribed in the original works. The ISOT dataset was divided into two parts using train_test_split into an 80-20 ratio. The testing data was left unseen and the performance metrics were derived using the unseen data. For DISCO, the

ISOT dataset is completely unseen and we thus test over the whole dataset. Below are the performance metrics of the three models:

Detection Models	ISOT		FakeNewsIndia	
	Precision	Recall	Precision	Recall
DISCO	-[1]	0.996	N/A[2]	0.877
NLP + GloVe	0.917	0.932	N/A[2]	0.894
FNDNet	0.995	0.996	N/A[2]	0.771

Table 4.2 Precision-Recall Scores

We can see that there is a clear difference in the performance of these models on data from the Global North and from the Global South, which supports our hypothesis that current disinformation detection models show a bias towards data originating from the Global North.

[1]Testing results only show the performance on 16754 entries from Fake.csv portion of the dataset. Since the percentage of Global South fake news is less than 5%, these articles are currently ignored in the comparison.

[2]Precision is not applicable as all articles belonging to this dataset are Fake.

Chapter 5

Conclusion and Further Work

The preliminary outcomes from testing diverse models on chosen datasets validate our hypothesis that existing disinformation detection systems exhibit better performance on data linked to the Global North. The apparent reason for this discrepancy may be the limited availability of fake news data from the Global South, potentially introducing bias. To address this, one potential avenue is to amass more fake news data from Global South regions, aiming to counterbalance the existing disparity.

The subsequent steps in our project will be dedicated to developing a disinformation detection system that tackles this bias. We plan to delve into concepts like pre-trained word embeddings, employing generative pre-training to introduce a broader world context, and exploring the integration of knowledge-graph-based methods with content-based detection.

The envisioned future work seeks to enhance the inclusivity of disinformation detection systems by mitigating the impact of bias introduced by the scarcity of data from Global South regions. The strategic collection of additional fake news data from these areas is anticipated to contribute to the creation of more equitable and globally applicable models. Our upcoming research endeavors will focus on implementing innovative approaches, including leveraging pre-trained word embeddings and generative pre-training techniques, as well as exploring the integration of knowledge-graph-based methods with content-based

detection. Through these efforts, we aim to advance the development of a disinformation detection system that excels on Global South data, fostering broader effectiveness and reducing biases in the field.

Appendix A

DISCO: Comprehensive and Explainable Disinformation Detection[1]

The authors claim that research regarding detecting disinformation is hindered by the heterogeneity of features, which can act as a camouflage for fake information. This is because words can be spun out of context using these features, effectively changing the semantics of the phrase. They suggest a computational framework that claims to model said heterogeneity using graph machine learning techniques and leverage it to solve the detection challenge. The model can also determine the importance of various words and sentences in the article using various graph augmentations (masking nodes and edges, etc.).

The suggested model uses pre-learned word embeddings and graph algorithms to determine the nature of the input content. The decision-making process of the proposed model can be split into four stages: word-graph generation from the input article, geometric feature extraction from the graph, neural detection, and ranking words based on their misleading degrees.

A.1 Word Graph

First, the input article is modeled by a word graph such that each word is represented by a node and two words that appear in the same k-word window are connected by a graph. Here, k is set to 3. Each node on the graph has a vector embedding derived from the large pre-trained model. The proposed model uses Google Word2Vec. A word graph G, containing n nodes is thus constructed, where n is the number of words in the article.

A.2 Geometric Feature Extraction

A word graph G is constructed with n nodes representing different words in an article. Node representations $\mathbf{h}_i$ are obtained through personalized PageRank vectors, encoding the stationary distribution of random walks from each node. The input node features $\mathbf{X}$ are transformed into specialized representations $\mathbf{h}_i$ using a geometric feature extraction process. The graph-level representation $\mathbf{u}$ is derived by aggregating word-level hidden representations. This method is effective by replacing traditional message-passing schemes and efficient, enabling fast updates with changing graph topology. The approach is applied in the DISCO system for semantic explanation.

A.3 Neural Detection

The DISCO model utilizes a cross-entropy loss function to evaluate its performance, comparing the representation vector $\mathbf{u}_j$ with label information $\mathbf{y}_j$ (e.g., indicating if a news article is fake or real). A neural network N transforms $\mathbf{u}_j$ into $\mathbf{z}_j$, and the geometric feature extraction allows for model-agnostic deployment of N, enabling the use of various neural networks. In the online demo, a simple 32×2 multi-layer perceptron (MLP) is employed for effective performance. Additionally, contextual multi-armed bandits are incorporated to address the exploitation-exploration dilemma.

A.4 Misleading Words

The paper introduces the concept of "misleading words" in the context of fake news detection. These words obscure the detection process, and their removal leads to more deterministic predictions. The DISCO system is employed to explain the misleading degree of each word by selectively masking nodes in the word graph G_j. The stationary distribution of seed nodes is tracked efficiently with topology changes, allowing for insight into factors influencing DISCO's predictions. The misleading degree is determined by the difference in correct prediction probabilities before and after masking. For instance, masking the word "recreational" in a news article increases the confidence of correct prediction, indicating its hindrance to DISCO's accuracy. The system ranks words based on their misleading degree, with negative values indicating words that assist correct predictions in the original article.

References

[1] D. Fu, Y. Ban, H. Tong, R. Maciejewski, and J. He, "Disco: Comprehensive and explainable disinformation detection," 2022.

[2] L. Yuan, H. Jiang, H. Shen, L. Shi, and N. Cheng, "Sustainable development of information dissemination: A review of current fake news detection research and practice," *Systems*, vol. 11, no. 9, p. 458, 2023.

[3] Z. Khanam, B. Alwasel, H. Sirafi, and M. Rashid, "Fake news detection using machine learning approaches," in *IOP conference series: materials science and engineering*, vol. 1099, no. 1. IOP Publishing, 2021, p. 012040.

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

[5] J. A. Nasir, O. S. Khan, and I. Varlamis, "Fake news detection: A hybrid cnn-rnn based deep learning approach," *International Journal of Information Management Data Insights*, vol. 1, no. 1, p. 100007, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667096820300070

[6] 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

[7] F. K. A. Salem, R. Al Feel, S. Elbassuoni, M. Jaber, and M. Farah, "Fa-kes: A fake news dataset around the syrian war," in *Proceedings of the international AAAI conference on web and social media*, vol. 13, 2019, pp. 573–582.

[8] 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.

[9] 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.

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