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

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.

1.1 Problem Statement

The problem statement can be divided into two sections. One part is to find and analyze the performance of current state-of-the-art disinformation detection techniques and models, with special emphasis on their performance based on the topic, location of origin, and the location of interest of the text in question. The goal is to test the idea that most disinformation detection techniques are tailored to the global North and do not perform equally as well on news from the global South, and to possibly identify the reasons for the same and find a way to mitigate said bias.

The second is to find datasets containing both real and fake news/information that can be used to test the said disinformation detection models. All the models must use the same set of articles for testing to ensure uniformity. However, additional features can be extracted and added or removed based on the requirements. The topic of the text can be extracted from the actual article itself based on keywords and style of writing. Each of the datasets will contain two features; the location of origin of the text and the location of interest of the text. This allows us to conduct various location-based analyses on these models.

Finally, create a disinformation detection system that works equally as well on data from both the global North and the global South, building on the ideas generated from our analyses.

1.2 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 takes a look at the work done so far on the project. It gives us insights into a few state-of-the-art techniques that are being

studied as a part of this project, as well as some of the difficulties faced while tackling the different tasks. Finally, chapter 3 talks about the immediate tasks and the ideal road map down the line.

Chapter 2

Work Done

2.1 Web Scrapping

Most datasets do not store the entirety of the articles as an entry, rather store the URL directing the user to the main article. A simple web scrapping script was written in order to extract the full article from these links whenever needed.

```
def extractArticles(inputFile, outputFile, range=-1):
  dataset_df = pd.read_csv(inputFile)
  df = pd.DataFrame(columns=['Link', 'Article'])
  count = 0

  for _, row in dataset_df.iterrows():
    if range != -1 and count == range:
      break
    link = str(row['news_url'])
    if link[0:4] != 'http':
      link = 'https://' + link
```

```
count = count + 1
try:
  response = urllib.request.urlopen(link, timeout=20)
  html = response.read()

  soup = BeautifulSoup(html, 'html.parser')

  for table in soup.find_all('table'):
    table.decompose()
  paragraphs = [data.get_text() for data in soup.find_all("p")]
  article = ' '.join(paragraphs)
  article = ' '.join(article.split())

except HTTPError as e:
    if e.code == 404:
        article = "Not found"
    else:
        article = f"Error: {e}"
except Exception as e:
    article = f"Error: {e}"
except HTTPError as e:
    if e.code == 404:
        article = "Not found"
    else:
        article = f"HTTP Error: {e}"
except ConnectionError as e:
    article = f"Connection Error: {e}"
except Exception as e:
    article = f"Error: {e}"
```

```
    df = df.append({'Link': link, 'Article': article}, ignore_index=True)

df.to_csv(outputFile, index=False)
```

The above function takes in the dataset in the form of a CSV file and writes the required output to the destination file, given by "outputFile". The range parameter was included for testing purposes, to ensure that the function runs only for a set number of entries defined by the same.

2.2 Data Acquisition

Multiple datasets were sourced from the internet. The credibility of these datasets was gauged by one of two factors: the entities responsible for its creation and/or the number of times the dataset was referenced by other similar academic literature. The following are the datasets that are currently in line to be used for testing the different disinformation detection techniques:

1. Politifact
2. Gossipcop
3. CoAid
4. FakeNewsIndia (PreCog, IIIT-H)
5. LIAR Dataset

A simple glimpse into these datasets shows us a hurdle that needs to be overcome: the lack of datasets concerning fake news in the global south.

2.3 Literature Review and Testing

The first task was to search for and collect working implementations of state-of-the-art disinformation detection techniques. The following papers were reviewed and subjected to testing:

2.3.1 Hyphen - Public Wisdom Matters! Discourse-Aware Hyperbolic Fourier Co-Attention for Social-Text Classification[1]

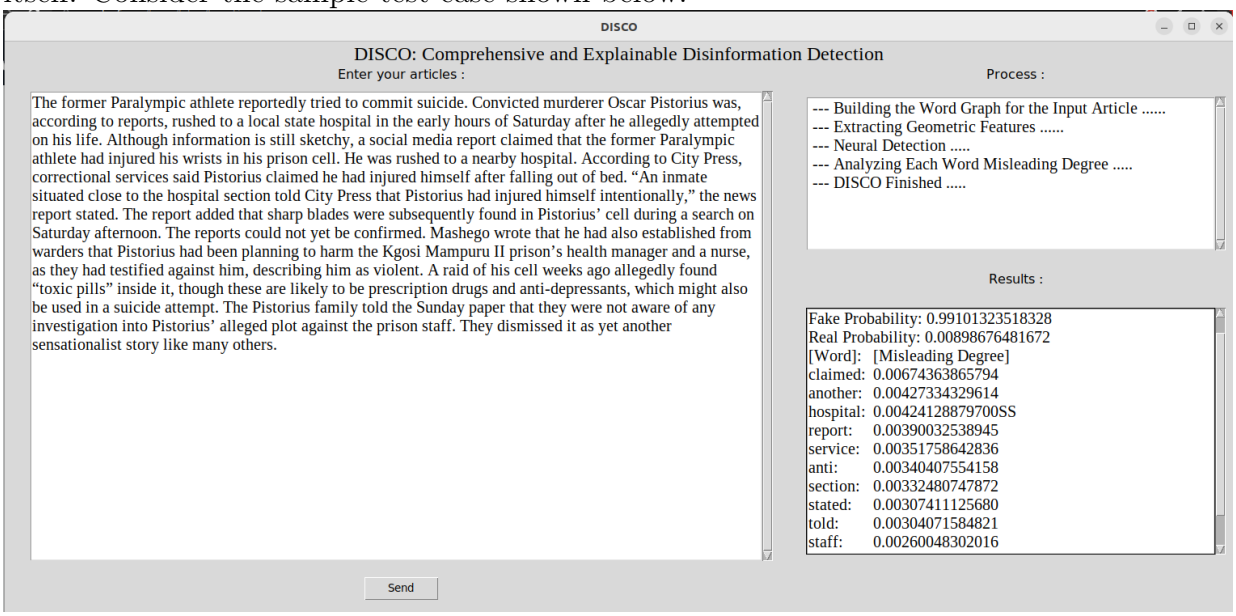
The "Hyphen" model is an innovative approach to classifying various forms of social media content, such as fake news, rumors, and sarcasm. It goes beyond traditional text analysis by incorporating user knowledge in the form of comments and replies to provide additional insights for classification. The model creates a graph-like abstraction of these comments and leverages this structure to determine the legitimacy of the associated headline. To be precise, Hyphen combines hyperbolic graph representation with a novel Fourier co-attention mechanism to capture the relationships between source posts and the associated comments. It has demonstrated state-of-the-art performance on ten benchmark datasets, making it a promising tool for understanding and categorizing complex social content.

However, several challenges have emerged in the testing of this model as a part of this project. One key limitation is the requirement for user comments as a dataset feature, which is often readily available on social media but significantly less common with traditional news articles. This lack of user comments in news datasets presents a considerable obstacle to testing the model's effectiveness. Moreover, building new datasets that incorporate user comments is a complex and resource-intensive task. Additionally, aligning the model with broader objectives, such as topic-based fake news analysis, is a challenge that requires further exploration. Keeping in mind all of these issues, discussion regarding "Hyphen" and testing have been tabled for later. While researching other techniques, solutions to overcome these roadblocks are being explored.

2.3.2 DISCO: Comprehensive and Explainable Disinformation Detection[2]

DISCO is a comprehensive system designed for the evaluation of the legitimacy of news articles. First, it constructs a word graph where words in the article become nodes, connected when they appear within a certain word proximity. These nodes are assigned individual feature vectors, enabling the analysis of word co-occurrence patterns. Geometric features are then extracted to understand the meaning of words in the context of their relationships within the graph, offering a comprehensive view of the article's content. Finally, a neural detection process calculates the probability of the article's authenticity and produces a credibility score. DISCO also identifies words that have the most impact on classifying the article as fake, generating a misleading degree ranking. This ranking serves to uncover words that obscure disinformation and provide insights into potential misclassifications. DISCO's approach aligns with research goals, including the analysis of writing style differences between real and fake news, cross-regional comparisons of fake news, and potentially extracting article topics.

Testing the model is a fairly simple task, as one needs only the article and the model itself. Consider the sample test case shown below.



We see that the model returns three key values: fake probability, real probability, and the list of words along with their misleading degree as explained in the previous paragraph. This is a very desirable outcome as we can see what factors contribute the most toward creating confusion and masking disinformation. Some logistical challenges continue to hinder progress in the required direction with this model. An immediate task to be tackled is to create a pipeline that can take multiple articles as input and store their outputs in a manner suitable for our purposes.

Chapter 3

Further Work

The upcoming work can be classed into two portions, immediate tasks, and future work. Immediate tasks deal more with the short-term goal of testing the DISCO model and analyzing the results. These tasks have already been discussed under the DISCO. The inferences of the same will be valuable observations and can serve as a benchmark for any model that we test after.

Long-term tasks can be divided into two portions, similar to the problem statement. Most disinformation detection datasets are geared towards the global North, hence it is imperative to find credible datasets that contain data regarding the global South. Furthermore, gathering fake news articles is difficult as the number of articles is much less compared to real news, and among them there exist articles that have been taken down or have broken URLs.

After gathering credible data, new test datasets must be created using these datasets. These new datasets will be used as the standard test datasets while comparing the performance of various disinformation detection models. The topic must remain consistent within each dataset, i.e., if the dataset is related to politics, then all the news articles in

this dataset must be only political. This ensures uniformity and allows us to analyze the performance of these models when dealing with individual topics. All the datasets must be annotated to include two new features, i.e., the location of origin of the article and the location of interest for the article. This task must be done manually, as we have to manually verify the accuracy of the assigned location.

On the other hand, more models that exhibit high-performance characteristics must be sought out. The more the number of models, the more diverse the study is about the underlying techniques. This allows us to possibly extract more information regarding why a bias exists between the global North and South, and also get better hints to tackle and overcome this problem. One short-term task in this segment is to find a solution to gather public discourse for mainstream news articles and headlines. If done, it enables the study to analyze and compare the performance of the Hyphen model with the rest.

Finally, by putting together the information we derive from the analysis of the different models, we aim to create a disinformation detection system that overcomes the bias between the global North and global South and performs equally well on fake news from both regions. This model will be trained using the newly created datasets from before. These datasets can be combined to form one large fake news repository, which can be made available to the public domain. Utmost care must be taken while combining these datasets because the datasets deal with different topics and any major imbalance within these topics would result in the model classifying the articles by placing a higher significance on these topics, which can result in undesirable behavior.

Aside from this, further work can be done to improve the working of the web scrapping tool. The current version, although functional in some cases, also picks up text from buttons, ads, and similar text containing regions from the given URLs. Removing this noise

is necessary as they are not a part of the article, but when included can contribute to the decision-making process of the model, leading to undesirable results.

3.1 References

[1] Karish Grover, S.M. Phaneendra Angara, Md. Shad Akhtar, Tanmoy Chakraborty, "Public Wisdom Matters! Discourse-Aware Hyperbolic Fourier Co-Attention for Social-Text Classification", 36th Conference on Neural Information Processing Systems (NeurIPS 2022)

[2] Dongqi Fu, Yikun Ban, Hanghang Tong, Ross Maciejewski, Jingrui He, "DISCO: Comprehensive and Explainable Disinformation Detection", Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, pp 4848–4852, 2022