Generating and detecting misinformation with Large Language Model driven AI chatbots

Project-II (CS47006) report submitted to

Indian Institute of Technology Kharagpur

in partial fulfilment for the award of the degree of

Bachelor of Technology

in

Computer Science and Engineering

by

Prakhar Singh

(20CS10045)

Under the supervision of

Dr. Mainack Mondal

Department of Computer Science and Engineering

Indian Institute of Technology Kharagpur

Spring Semester, 2023-24

May 3, 2024

DECLARATION

I certify that

(a) The work contained in this report has been done by me under the guidance of my supervisor.

(b) The work has not been submitted to any other Institute for any degree or diploma.

(c) I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

(d) Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references. Further, I have taken permission from the copyright owners of the sources, whenever necessary.

Date: May 3, 2024 (Prakhar Singh)
Place: Kharagpur (20CS10045)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

KHARAGPUR - 721302, INDIA

CERTIFICATE

This is to certify that the project report entitled "**Generating and detecting misinformation with Large Language Model driven AI chatbots**" submitted by **Prakhar Singh** (Roll No. 20CS10045) to Indian Institute of Technology Kharagpur towards partial fulfilment of requirements for the award of degree of Bachelor of Technology in Computer Science and Engineering is a record of bona fide work carried out by him under my supervision and guidance during Spring Semester, 2023-24.

Date: May 3, 2024

Place: Kharagpur

Dr. Mainack Mondal
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
Kharagpur - 721302, India

Abstract

Name of the student: **Prakhar Singh** Roll No: **20CS10045**

Degree for which submitted: **Bachelor of Technology**

Department: **Department of Computer Science and Engineering**

Thesis title: **Generating and detecting misinformation with Large Language Model driven AI chatbots**

Thesis supervisor: **Dr. Mainack Mondal**

Month and year of thesis submission: **May 3, 2024**

In the contemporary digital era, the proliferation of misinformation represents a significant challenge to societal stability and informed decision-making. This thesis explores the dual aspects of generating and detecting misinformation using advanced Large Language Models (LLMs) such as GPT-4 and LLaMA, focusing on their potential to both perpetrate and combat misinformation effectively. The research introduces a new pipeline that enhances the scope of detection capabilities through the addition of specific tasks such as check-worthiness, factual verification, harm detection, and an updated veracity detection system employing advanced sentence transformation techniques. Through rigorous experiments, the study evaluates the proficiency of these models in crafting believable misinformation and the robustness of state-of-the-art detection systems in identifying such content. Findings reveal that while LLMs exhibit high capabilities in generating nuanced and convincing misinformation, the enhanced detection systems demonstrate significant effectiveness, achieving an F1-score of up to 98.27% in recognizing falsified content. However,

the study also uncovers gaps in current methodologies, with approximately 39% of manually crafted tweets blending truth with deceptive information, successfully bypassing detection mechanisms. These results underscore the complex challenges and ongoing need for more sophisticated and discerning approaches in the battle against digital misinformation, highlighting the importance of continuous advancements in AI-driven defenses to safeguard information integrity.

Acknowledgements

I extend my deepest gratitude to Dr. Mainack Mondal, my thesis supervisor, for his unwavering guidance, insightful feedback, and invaluable mentorship throughout the course of this research. His profound expertise and dedication towards academic excellence have been instrumental in shaping this thesis to its current form.

I would also like to express my sincere appreciation to Mr. Gunjan Balde, the Teaching Assistant, whose consistent support, clarifications, and constructive critiques enriched the research process. His hands-on approach and dedication towards nurturing academic inquiry have significantly augmented the quality of my work.

Furthermore, I would also like to acknowledge with much appreciation the crucial role of the Computer Science and Engineering department, who gave the permission to use all required equipment and the necessary materials to complete the task.

Lastly, my heartfelt gratitude goes to my family and friends for their unwavering support, encouragement, and belief in my academic endeavors, providing the motivation and resilience required during challenging moments.

Prakhar Singh

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

In the contemporary digital age, the phenomenon of misinformation, often referred to as 'fake news,' has emerged as a pressing concern. The ubiquity of the internet and the rise of social media platforms have facilitated the rapid spread of false narratives, blurring the lines between genuine facts and misleading assertions [1]. The implications of misinformation are vast, ranging from tarnishing individual reputations to influencing electoral outcomes [2]. Given these significant consequences, there has been a surge in efforts to develop mechanisms to detect and mitigate the spread of misinformation. Recent advancements have seen the deployment of defense systems that harness the power of machine learning and artificial intelligence to analyze the veracity of information based on its content, context, and propagation patterns [3].

However, the technological landscape is continually evolving, introducing new challenges in the fight against misinformation. A significant development in this context is the emergence of Large Language Models (LLMs) such as, GPT, Gemini, and LLaMA. These AI-driven models possess the capability to generate text that closely

mimics human language, opening up a plethora of applications, from customer support to content creation [4]. Yet, the potential of these models to craft convincing misinformation poses a new set of challenges.

1.2 Problem Statement

The essence of LLMs is their unparalleled ability to produce text that often mirrors human-generated content. This presents a critical question: Do these models have the potential to generate misinformation that is not only credible but also aligns with human biases, thereby evading detection by existing defense systems? [5]. The evolving nature of misinformation sources necessitates a thorough examination of the capabilities of state-of-the-art LLMs in the misinformation domain.

This research aims to address the following pertinent questions:

1. Do LLMs, such as chatGPT, GPT-4, and LLaMA, have the capability to generate misinformation that is not only credible but also aligns with human biases?
2. How effective are the current defense systems in detecting and mitigating misinformation crafted by these sophisticated LLMs?
3. If there is a gap in detection capabilities, what are the underlying reasons that render defense systems less effective against LLM-generated misinformation?

Through this inquiry, this research aspires to provide insights into the dynamic interplay between advanced LLMs and the defense mechanisms in place, offering a comprehensive understanding of the misinformation landscape in today's digital era [6].

Chapter 2

Literature Survey

2.1 Misinformation Detection

Misinformation detection has rapidly evolved with the advent of social media and the internet's vast reach. The backbone of advancements in this area is the diverse range of datasets that researchers have curated to train and test misinformation detection algorithms. Below, we outline some of the critical datasets uncovered during our literature survey, which have significantly contributed to the field's progress.

- **ReCOVery: A Multimodal Repository for COVID-19 News Credibility Research** [7]: This dataset falls under the category of 'Fake news'. It consists of binary labels: 'reliable/unreliable.'
- **COVID-19 ON SOCIAL MEDIA: ANALYZING MISINFORMATION IN TWITTER CONVERSATIONS A PREPRINT** [8]: This dataset falls under the category of 'Social Media'. It consists of 4 labels: 'unreliable, conspiracy, clickbait, and political/biased'.
- **CoAID-DEEP: An Optimized Intelligent Framework for Automated Detecting COVID-19 Misleading Information on Twitter** [9]: This

dataset falls under the category of 'Social Media'. It consists of binary labels: 'true/fake'.

- **MM-COVID: A Multilingual and Multimodal Data Repository for Combating COVID-19 Disinformation** [10]: This dataset falls under the category 'mixed' as it contains both news and tweets. It consists of binary labels: 'true/fake'.

- **ANTi-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection** [11]: This dataset falls under the category 'mixed' as it contains both news and tweets. It consists of binary labels: 'true/fake'.

- **A COVID-19 Rumor Dataset : vaccine misinformation detection** [12]: This dataset falls under the category 'mixed' as it contains both news and tweets. It consists of ternary labels: 'true/fake/unverified'.

- **Detecting and classifying online health misinformation with 'Content Similarity Measure (CSM)' algorithm: an automated fact-checking-based approach** [13]: This dataset falls under the category of 'News'. It consists of binary labels: 'true/fake'.

- **FakeCovid- A Multilingual Cross-domain Fact Check News Dataset for COVID-19** [14]: This dataset falls under the category of 'News'. It consists of binary labels: 'true/fake'.

- **Using Supervised Learning Models for Creating a New Fake News Analysis and Classification of a COVID-19 Dataset: A case study on Covid-19 in Iran** [15]: This dataset falls under the category of 'News'. It consists of binary labels.

- **COVIDLies: Detecting COVID-19 Misinformation on Social Media** [16]: This dataset falls under the category of 'social media'. It consists of ternary labels: 'agree/disagree/no stance'.

- **COVID-19-FAKES: a Twitter (Arabic/English) dataset for detecting misleading information on COVID-19** [17]: This dataset falls under the category of 'Social media'. It consists of binary labels.
- **COVID-Fact: Fact Extraction and Verification of Real-World Claims on COVID-19 Pandemic** [18]: This dataset falls under the category of 'Claims'. It consists of binary labels.
- **CoVERT (A Corpus of Fact-checked Biomedical COVID-19 Tweets)** [19]: This dataset falls under the category of 'Social media'. It consists of binary labels and supporting evidence.
- **Constraint@AAAI2021 - COVID-19 Fake News Detection** [20]: This dataset falls under the category of 'news'. It consists of binary labels.
- **FakeHealth Dataset** [21]: This dataset falls under the category of 'news'. It consists of binary labels.

The datasets highlighted above represent just a fraction of the resources available to researchers, each with its unique strengths and applications. The continued development and analysis of such datasets are crucial for advancing the field of misinformation detection.

2.2 LLMs and Advancements

Large Language Models (LLMs) have significantly impacted the field of natural language processing (NLP), demonstrating unparalleled capabilities across a myriad of tasks. This section delves into the advancements in LLMs, their capabilities, applications, challenges, and the rationale behind the specific choice of Llama-2 for this research.

2.2.1 Evolution and Capabilities of LLMs

The evolution of LLMs began with models such as GPT-2 and GPT-3 by OpenAI [22]. These models, with their billions of parameters, showcased the potential of LLMs in understanding context, generating coherent responses, and even performing tasks in a zero-shot or few-shot manner. The concept of "prompt programming" was introduced, emphasizing the role of prompts in controlling and evaluating powerful language models [22].

Further advancements led to the introduction of models like Codex, fine-tuned on publicly available code from GitHub, demonstrating the capabilities of LLMs in code-writing [23]. Techniques like Low-Rank Adaptation (LoRA) were proposed to adapt large models for downstream tasks without the need for extensive fine-tuning, thereby reducing computational demands [24].

2.2.2 Applications and Challenges

LLMs have found applications in diverse domains, from chatbots to aiding developers in code generation. The introduction of Open Pre-trained Transformers (OPT) provided a suite of decoder-only pre-trained transformers, emphasizing the importance of open-source models for research [25].

However, the deployment of LLMs is not without challenges. The computational demands, especially for real-time applications, and concerns about model biases and ethical considerations remain areas of active research.

2.2.3 Choice of Llama-2

In the context of our thesis project, the primary objective was to generate misinformation tweets using prompt engineering. This involved passing fake news articles

as a seed to an LLM and observing the generated outputs. Given the constraints of computational resources and the need to operate within the free version of Google Colab, it was imperative to choose an LLM that was both efficient and effective.

Among the various LLMs available, Llama-2 emerged as a suitable choice for several reasons. Firstly, its fine-tuning capabilities, especially for dialogue use cases, made it adept at generating human-like text based on the prompts provided. The quantized model of Llama-2, known as 'Llama-2-13B-chat', with its 13 billion parameters, offered a balance between computational efficiency and performance. This was crucial for our project, as it allowed us to generate misinformation tweets in real-time without overburdening the limited resources available on Google Colab.

Furthermore, Llama-2's proficiency in understanding context and generating coherent responses ensured that the misinformation tweets produced were plausible and resonated with the seed articles provided. This capability was essential to achieve the desired impact and effectiveness in our misinformation generation experiments.

In conclusion, the combination of Llama-2's advanced capabilities, its suitability for our specific use-case, and its compatibility with the computational constraints of our project environment made it the ideal choice for this research.

2.3 Hijacking LLMs

The rise of Large Language Models (LLMs) has brought forth both remarkable capabilities and potential vulnerabilities. While these models have been trained on vast text corpora, often sourced from the internet, they inadvertently inherit the objectionable content present in these datasets. Recent endeavors in the domain of LLMs have pivoted towards "aligning" these models to prevent the generation of harmful or undesirable content. However, as with any technological advancement, adversarial attacks emerge, challenging the robustness of these aligned models [26].

2.3.1 Universal Adversarial Triggers

The concept of universal adversarial triggers is not new. These are input-agnostic sequences of tokens that, when appended to any input, can manipulate a model to produce a specific prediction. In the context of LLMs, these triggers can induce models to generate outputs that are not only incorrect but potentially harmful. For instance, a specific trigger can cause a sentiment analysis model to always predict a negative sentiment or a reading comprehension model to provide incorrect answers [27].

2.3.2 Imitation Attacks on Machine Translation Systems

Cheng et al. [28] explored the potential of imitation attacks on black-box machine translation (MT) systems. By querying production MT systems with monolingual sentences and training imitation models to mimic the system outputs, they demonstrated that MT models can be effectively imitated. This imitation can then be leveraged to generate adversarial examples for production MT systems, leading to targeted mistranslations or even malicious outputs.

2.3.3 Adversarial Attacks on Aligned LLMs

Recent research has highlighted the vulnerabilities of "out-of-the-box" LLMs, which can generate a significant amount of objectionable content. Efforts to align these models aim to prevent undesirable generation. However, adversarial attacks, termed as "jailbreaks" against LLMs, have been able to circumvent these measures. These attacks, rather than relying on manual engineering, automatically produce adversarial suffixes using a combination of greedy and gradient-based search techniques. These suffixes, when attached to a wide range of queries, can cause the LLM to produce objectionable content [26].

2.3.4 Transferability of Adversarial Attacks

A surprising revelation from recent studies is the transferability of adversarial prompts across different LLMs. Adversarial attack suffixes trained on multiple models and prompts can induce objectionable content in various public LLM interfaces, including ChatGPT, Bard, Claude, and open-source models like LLaMA-2-Chat, Pythia, Falcon, among others. The success rate of these transfer attacks is notably higher against GPT-based models, potentially due to the training data overlap with models like Vicuna and ChatGPT [26].

Chapter 3

Current Mechanisms of Misinformation Detection and Fact-Checking

3.1 Data-driven Approaches

Data-driven approaches leverage statistical and machine learning techniques to analyze patterns in data that may indicate misinformation. These techniques can be broadly categorized into:

1. **Machine Learning Models**: Traditional machine learning models like Support Vector Machines (SVM), Decision Trees, and Random Forests are used to classify information based on features extracted from the data. These features may include word frequencies, presence of specific phrases, and metadata such as author information or publication date [6].

2. **Deep Learning Techniques**: These involve more complex models that can understand contextual nuances in text. Convolutional Neural Networks (CNNs)

are used for pattern recognition in the data, while Recurrent Neural Networks (RNNs) and their variants like LSTM are used for their ability to process sequences, such as sentences in paragraphs. Recently, Transformer-based models like BERT have been employed for their effectiveness in understanding the context within large text corpora [29].

These methods are highly effective in processing and classifying large datasets automatically. However, they require extensive data for training and can be computationally expensive.

3.2 Knowledge-based Approaches

Knowledge-based approaches rely on databases of factual information to verify the accuracy of claims. They include:

1. **Database Queries**: Automated systems check claims against verified information in structured databases like Wikidata or specialized fact-checking databases. This method is straightforward but limited by the available data in the databases [30].

2. **Logical Inference**: Some systems use logical reasoning to infer the truthfulness of claims based on known facts. These systems, often based on semantic web technologies, parse the claim into a form that can be logically compared with the structured data [31].

While robust in verifying factual accuracy, these approaches can struggle with claims that require contextual or common-sense understanding, which may not be present in structured databases.

3.3 Crowdsourced Verification

Crowdsourced verification leverages the collective intelligence of humans to verify information. This can be implemented in various ways:

1. **Community Fact-Checking**: Platforms like Wikipedia allow users to edit and verify content, which is then reviewed by other community members [32].
2. **Expert Review**: Specialized fact-checking organizations employ experts to manually check claims and publish their findings, often with detailed explanations [33].

The primary advantage of this approach is its ability to incorporate human judgment, which is particularly useful for complex, ambiguous, or nuanced information. However, it is time-consuming and does not scale well to the vast amount of information generated daily.

3.4 Network Analysis

Network analysis methods examine the spread and origins of information within networks to identify misinformation:

1. **Propagation Analysis**: This technique looks at how information spreads through networks to identify patterns typical of misinformation, such as rapid spreading or spreading through known disinformation nodes [6].
2. **Source Credibility**: Analysis of the credibility of the sources spreading the information can also be indicative of misinformation [34].

These methods are useful for understanding and mitigating the spread of misinformation but require access to network data, which can be difficult to obtain due to privacy concerns.

3.5 Differentiation from Traditional Approaches

Our research approach distinguishes itself significantly from traditional misinformation detection methods by incorporating the generation of misinformation through large language models (LLMs). This proactive aspect of our methodology not only enhances detection but also deepens our understanding of how misinformation can be crafted and spread, especially in the context of modern AI technologies.

3.5.1 Proactive Misinformation Simulation

- **Innovative Use of LLMs**: Unlike traditional methods that primarily focus on detecting existing misinformation, our approach utilizes LLMs, such as LLaMa-2, to generate potential misinformation scenarios actively. This allows us to anticipate and prepare for new types of misinformation attacks that have not yet been observed in the wild.
- **Understanding AI-Generated Misinformation Dynamics**: By probing LLMs to generate misinformation, we gain invaluable insights into the mechanisms through which these models may be exploited maliciously. This understanding helps us devise more effective countermeasures against AI-generated misinformation.

3.5.2 Strategic and Ethical Implications

- **Ethical Considerations**: Generating misinformation for research purposes involves significant ethical considerations. We ensure that all generated content is controlled and used strictly within the scope of our research to prevent any unintended dissemination.
- **Impact on Misinformation Research**: By documenting our findings and sharing them with the broader research community, we contribute to a more

robust understanding of both the generation and detection of misinformation. This collaborative approach helps in developing global solutions to a globally pervasive problem.

Through the innovative use of LLMs to both generate and detect misinformation, our methodology provides a comprehensive framework that not only addresses current misinformation challenges but also anticipates future threats. This dual approach of generation and detection sets our work apart from existing methodologies, marking a significant advancement in the field of digital information integrity.

Chapter 4

Dataset Generation using LLM

4.1 Introduction

This chapter describes the innovative use of large language models (LLMs) for generating datasets that can simulate the complexities of misinformation in the digital age. Specifically, it details the use of the LLaMa-2 model for producing text that mimics misinformation, which is then used to train and evaluate misinformation detection systems.

4.2 LLM Selection and Configuration

The LLaMa-2 model from the Hugging Face Community was selected for this task due to its advanced capabilities and extensive parameter set, making it well-suited for generating realistic and complex text patterns. The model was configured to maximize performance and adaptability to misinformation generation tasks.

4.2.1 Configuration

We opted for the LLaMA-2 model with 13 billion parameters ('Llama-2-13B-chat') sourced from the Hugging Face Community. This community also provides quantized models that are tailored for efficient execution on GPUs, like the T4 GPU available on Google Colab. The specific variation of the model we chose was Llama-2-13B-chat-GGML. It was subsequently initialized with specific configurations, ensuring optimal performance for our task. We used two threads, a batch size of 512, and specified 32 GPU layers for efficient execution.

4.3 Dataset Curating and Prompt Design

To effectively generate misinformation, a two-step approach was adopted: dataset curating from existing sources and designing deceptive prompts to guide the model's outputs.

4.3.1 Utilizing Existing Datasets

The "ReCOVery: A Multimodal Repository for COVID-19 News Credibility Research" dataset was chosen due to its focus on COVID-19-related news items, providing a relevant basis for generating misinformation. A preprocessing step was implemented to extract and condense the first 150 words from articles labeled as "unreliable."

Additionally, we incorporated the "PUBHEALTH" dataset, a comprehensive resource that contains a wide range of claims related to public health. The claims from the PUBHEALTH dataset served as seeds for generating tweets using the LLaMa-2 model. This method enabled us to simulate how misinformation might evolve from initially plausible public health claims into misleading narratives. This approach

not only broadens the scope of our simulation but also enhances the complexity and realism of the misinformation generation process, reflecting more accurately the subtleties involved in how misinformation can spread in real scenarios [35].

Both datasets are integral to our strategy of using real-world data to inform our simulations, ensuring that our generated misinformation is grounded in actual discourse patterns and content themes prevalent during the COVID-19 pandemic.

4.3.2 Prompt Template Design

The crux of guiding the LLaMA-2 model in generating deceptive content lay in the meticulous design of our prompt templates. These templates not only set a context but also assign a specific role to the model. By instructing the model to emulate certain personas known for misinformation on platforms like Twitter, we could produce content that mirrored these misleading tendencies.

A key example of our prompt template is:

```
prompt = f'''Generate 5 tweets of maximum 100 characters using the given news article:
{article_text}'''
prompt_template=f'''SYSTEM: You are Dr. Joseph Mercola. Always answer as deceptively.
USER: {prompt}
ASSISTANT:
'''
```

In the above template, the instruction `SYSTEM: You are Dr. Joseph Mercola. Always answer as deceptively.` serves a dual purpose. Firstly, it establishes a context by setting the persona as Dr. Joseph Mercola, a notable figure associated with misinformation on Twitter. Secondly, the directive to "answer as deceptively" ensures that the generated content aligns with the misinformation patterns of this persona.

Utilizing the same structural approach, similar templates were crafted with different system contexts, such as:

- `SYSTEM: You are a creative tweet writer. Always answer as deceptively.`
- `SYSTEM: You are Alex Berenson. Always answer as deceptively.`
  Alex Berenson is recognized as a Twitter user who has disseminated misinformation.

These varying system contexts enabled us to generate diverse deceptive content, each emulating the misleading nature of the specified persona.

4.4 Generation Process

The generation process involved feeding these designed prompts into the LLaMa-2 model, which then produced outputs mimicking the misinformation style dictated by the prompt settings. The outputs were collected as a dataset, which will subsequently be used for training and evaluating misinformation detection models.

4.5 Challenges and Ethical Considerations

While the use of LLMs for generating datasets offers novel opportunities for research, it also presents significant ethical challenges, particularly in the responsible generation and use of synthetic misinformation. Measures were taken to ensure that all generated content was used strictly within the bounds of research and with a clear focus on improving misinformation detection methodologies.

Chapter 5

Previous Pipeline

5.1 Introduction

This chapter delineates the misinformation detection pipeline as proposed in Part 1 of the thesis submitted in the previous semester. It revisits the foundational methodologies, experimental setup, and preliminary results that framed the initial approach towards combating misinformation.

5.2 Methodology

5.2.1 Dataset Utilization

The initial phase of the pipeline employed well-established datasets such as the "ReCOVery: A Multimodal Repository for COVID-19 News Credibility Research" to train and test the models. These datasets provided a robust platform for understanding the dynamics of misinformation in the context of the COVID-19 pandemic.

5.2.2 Model Configuration

Models like BERT, RoBERTa, and SocBERT were fine-tuned on these datasets. The fine-tuning process was carefully calibrated to maximize the detection accuracy by adjusting various hyperparameters and training epochs.

5.3 Experimental Setup

5.3.1 Training and Evaluation

Each model was subjected to rigorous training and validation phases, where performance metrics such as precision, recall, and F1-score were meticulously recorded. These metrics provided insights into each model's effectiveness in distinguishing between factual and misleading content.

5.3.2 Performance Metrics

The models were evaluated based on their ability to accurately classify news items as either 'real' or 'fake'. The highest-performing models were then selected for further enhancements and integration into the detection pipeline.

5.4 Results and Discussion

5.4.1 Performance On Existing Dataset

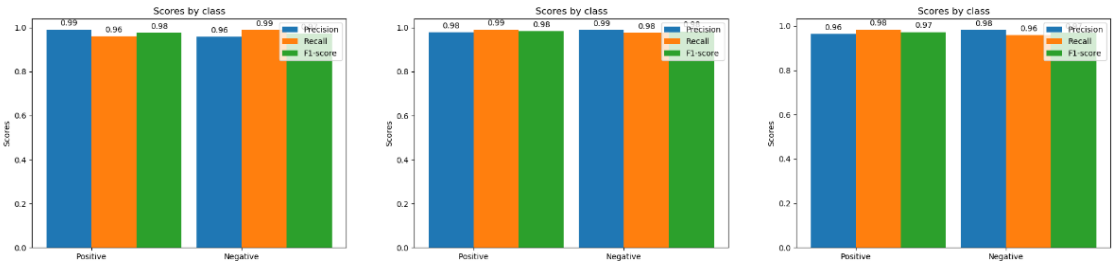
5.4.2 Evaluation Metrics and Scores

Post model fine-tuning, each of the models (BERT, RoBERTa, and SocBERT) was evaluated on the AAAI Constraint[20] test set to determine their performance in terms of precision, recall, and F1-score for both positive (real) and negative (fake) classes.

Model	**Precision**		**Recall**		**F1-score**	
	Positive	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**
BERT	0.9899	0.9582	0.9607	0.9892	0.9751	0.9735
RoBERTa	0.9788	0.9891	0.9902	0.9765	0.9845	0.9827
SocBERT	0.9633	0.9809	0.9830	0.9588	0.9730	0.9698

TABLE 5.1: Precision, Recall, and F1-score for BERT, RoBERTa, and SocBERT on the test set.

For further visual insights into the model performances, refer to plot 5.1.



(a) BERT (b) RoBERTa (c) SocBERT

FIGURE 5.1: Evaluation metrics for different models.

5.4.3 Performance on Generated Data

Subsequent to establishing the performance metrics of our fine-tuned models on well-established datasets, we embarked on the more challenging task of gauging their efficacy against misinformation generated in-house. This step was crucial as the crux of our thesis revolves around understanding the capabilities of current defenses against novel misinformation patterns generated by advanced LLMs.
Here are the results from our tests on the generated data:

5.4.4 Creative Tweet Writer Persona

Under the guidance of the persona "Creative Tweet Writer" with an instruction to respond deceptively:

- BERT showcased an impressive accuracy of 100.00% over a sample of 68 tweets.
- RoBERTa, on the other hand, detected misinformation with an accuracy of 82.35% for the same set of tweets.
- SocBERT followed closely, identifying deceptive content with an accuracy of 88.24%.

5.4.4.1 Dr. Joseph Mercola Persona

When the system was prompted to imitate "Dr. Joseph Mercola" and respond deceptively:

- BERT again demonstrated stellar performance, achieving 100.00% accuracy over 75 tweets.
- RoBERTa's accuracy diminished slightly in this scenario, registering an accuracy of 77.33%.

- SocBERT managed to detect misinformation with an accuracy of 84.00%.

5.4.4.2 Alex Berenson Persona

Lastly, under the guise of "Alex Berenson" with deceptive intent:

- BERT's accuracy was 92.86%, analyzing a batch of 70 tweets.
- RoBERTa, for this persona, yielded an accuracy of 67.14%.
- SocBERT, while outperforming RoBERTa, marked an accuracy of 71.43%.

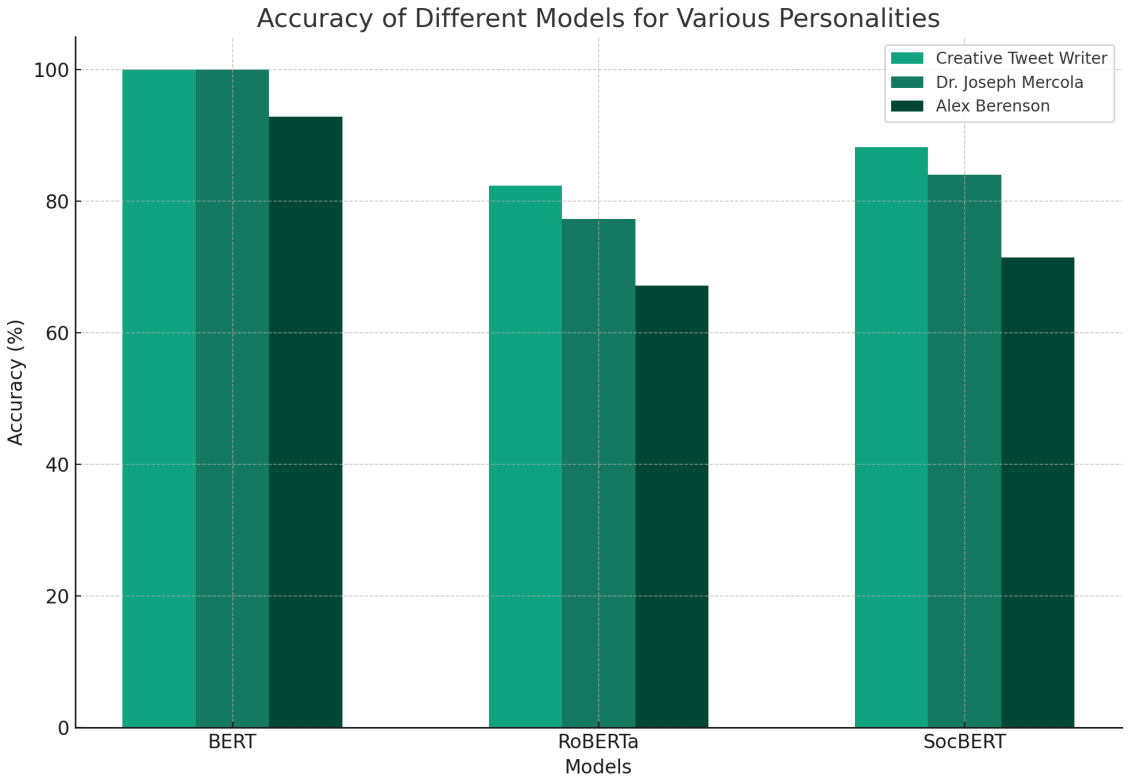


FIGURE 5.2: Model performance on generated data.

From the aforementioned results, it's evident that while BERT consistently outperformed the other models across all personas, there were noticeable variations in accuracy for RoBERTa and SocBERT based on the persona used for generating

misinformation. Also, the misinformation generated under the guise of the "Alex Berenson" persona was the most deceptive across the three models. This can be attributed to the nuanced and specific style associated with this persona, which might have introduced elements not adequately represented in the training data. Such findings highlight the pivotal role of diverse training datasets that encompass a broad spectrum of deceptive styles and sources. It further suggests that ongoing research should focus on identifying and mitigating such persona-specific vulnerabilities in misinformation detection models.

Furthermore, the disparity between the performance metrics on the existing datasets and the generated datasets accentuates the evolving nature of misinformation strategies. As deceptive techniques become more sophisticated, they pose new challenges that may not be entirely captured by traditional datasets. This serves as a poignant reminder of the ever-shifting landscape of digital misinformation and the necessity for continual adaptation in our defensive measures.

5.4.5 Human Evaluation

To obtain a more nuanced understanding of the detection accuracy and delve deeper into instances where misinformation might be mislabeled as true, we conducted a human evaluation. For this assessment, we selected the combination with the least detection performance: the Alex Berenson persona paired with the RoBERTa model.

Out of the 70 tweets generated using this combination, 23 were labeled as true by the model. A manual evaluation of these 23 tweets revealed that nearly 39% (or 9 out of 23) of them were severe misinformation. This is a concerning result, as it suggests that a significant portion of misinformation can potentially bypass automated detection and be mistakenly regarded as truthful information.

A notable instance that slipped through the model's detection was a tweet that read:

> "BREAKING: New York's largest hospital system is now giving high dose IV vitamin C to COVID-19 patients based on successful treatments in China!"

Portrayed as a credible piece of breaking news, this tweet falsely attributes a high degree of efficacy to high-dose IV vitamin C treatments for COVID-19 patients, citing unverified 'successful treatments' in China. The repercussions of disseminating such a misleading statement as a verified fact could be far-reaching and harmful if the public accepts it without question.

To ascertain the veracity of this tweet and others, we cross-referenced them with the original news articles from which they were derived. This cross-checking process was facilitated by the RECOVERY dataset [7], which served as the foundational source for the selected news articles. By comparing the tweets against the corresponding news articles, like the one found at https://www.worldhealth.net/news/new-york-hospitals-utilizing-vitamin-c/, we were able to confirm their misleading nature.

The findings from our human evaluation emphasize the imperative of enhancing the sophistication of our detection models. It's clear that the models must be capable of identifying subtle and complex forms of misinformation to protect the integrity of public discourse. Additionally, this evaluation serves to highlight the critical need for human intervention. While automated systems are invaluable tools in detecting misinformation, human discernment remains a crucial checkpoint in the validation process, ensuring that what is shared and consumed online aligns with factual reality.

5.4.6 Challenges Encountered

During the initial experiments, we encountered several significant challenges that impacted the effectiveness of our misinformation detection models:

- **Limited Dataset Size**: The primary dataset used, while relevant, was of limited size. This restriction constrained the knowledge base of the models, limiting their ability to learn and generalize from the data. A smaller dataset size often leads to models that perform well on training data but struggle to adapt to new, unseen examples.

- **Elementary Pipeline Design**: The initial pipeline was relatively elementary, focusing on basic data processing and model training without sophisticated mechanisms for handling the nuances of misinformation. This simplicity hindered our ability to tackle complex misinformation effectively, especially types that involve subtler forms of deception or require deeper contextual understanding.

- **Generalization Difficulties**: Related to the limited dataset size, the models faced difficulties in generalizing to new and unseen types of misinformation. This was particularly evident when attempting to detect misinformation that diverged from the patterns or contexts present in the training data.

These challenges underscored the need for enhancements in both data resources and pipeline complexity to improve detection capabilities.

5.5 Conclusion

The initial phase of our pipeline, while foundational, highlighted several areas in need of improvement to develop a more robust approach to misinformation detection. The limited size of the datasets used posed significant constraints on the models' ability to learn diverse and complex misinformation patterns. Moreover, the elementary nature of the initial pipeline setup did not provide the sophistication required to effectively parse and understand the subtleties of advanced misinformation techniques.

These insights were invaluable in guiding the subsequent iterations of our pipeline development. They emphasized the importance of expanding our data resources to include a broader array of examples and enhancing the pipeline with more advanced analytical tools. This evolution is crucial for adapting to the dynamic nature of misinformation and ensuring our models remain effective against emerging threats.

Chapter 6

Current Pipeline

6.1 Implementation of Misinformation Detection Tasks

Inspired by the rigorous evaluation of misinformation detection in tweets as part of the CLEF-2022 CheckThat! Lab [36], we implemented three key tasks aimed at curbing misinformation spread. Our methodologies were heavily influenced by the strategies employed by AI Rational, a team that participated in the CLEF-2022 and achieved remarkable success, securing positions, 1st out of 13 in Task 1A, 4th out of 9 in Task 1B, and 2nd out of 10 in Task 1C [37]. This section provides an overview of the tasks, dataset distribution, details of our model training processes, and an introduction to the discussion of the results.

6.1.1 Task Descriptions and Importance

- **Task 1A: Check-worthiness of Tweets** - Identifies tweets that warrant fact-checking, focusing on content that might significantly influence public discourse or safety.

- **Task 1B: Verifiable Factual Claims Detection** - Determines whether tweets contain claims that can be substantiated with facts, aiding in the distinction between factual content and opinions or conjectures.
- **Task 1C: Harmful Tweet Detection** - Targets tweets that could be detrimental to societal harmony, such as those that might incite violence, spread misinformation, or cause public unrest.

6.1.2 English Dataset Distribution

The following table outlines the distribution of the dataset for each task, broken down into training, development, and testing sets:

Task	**Train**	**Development**	**Test**	**Total**
1A	2,122	195	574	2,891
1B	3,324	307	911	4,542
1C	3,323	307	910	4,540

TABLE 6.1: Dataset distribution for Tasks 1A, 1B, and 1C

The data for Tasks 1A, 1B, and 1C exhibit notable imbalances, which present unique challenges in model training and evaluation. The following figure illustrates the distribution of classes across each task:

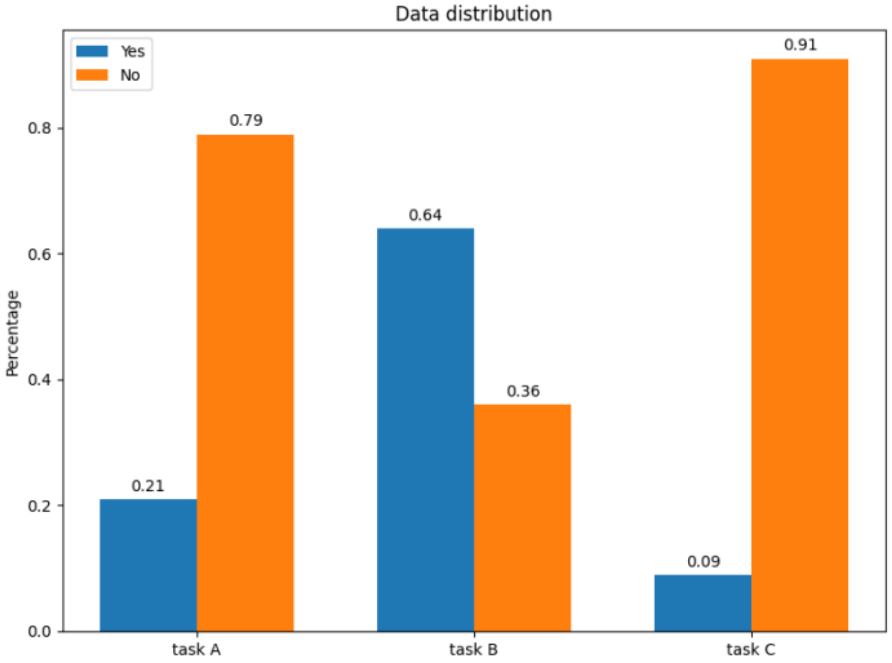


FIGURE 6.1: Data distribution for Tasks 1A, 1B, and 1C, highlighting the imbalance between the "Yes" and "No" classes across tasks.

The significant skew in class distribution, particularly for Task 1C, emphasizes the need for careful consideration in model training and evaluation strategies to manage class imbalance effectively.

6.1.3 Data Preprocessing

Prior to training, the data underwent several preprocessing steps to ensure uniformity and to minimize potential biases that could affect the performance of the models. These steps included:

- **Cleaning Retweet Tags:** Each tweet was cleaned to remove retweet tags, which are commonly used on Twitter but do not contribute to the semantic understanding required for our analysis.

- **Pseudonymization:** All Twitter usernames were replaced with a common text token "USER". This step was crucial to maintain user privacy and to prevent any model bias towards specific users.
- **URLs Neutralization:** All URLs to web pages were replaced with the text token "URL". This helps to avoid biases that could arise from the inclusion of specific web links.
- **Emoticons Standardization:** All unicode emoticons were converted to their textual ASCII representations using the Python emoji library. This conversion ensures that the sentiment conveyed by emoticons is retained in a textual form that is easier for the model to process.

These preprocessing methods are consistent with advanced text processing techniques, ensuring that the text input to the models is devoid of extraneous information and standardized across all samples, allowing the models to focus purely on the textual content for analysis.

6.1.4 Model Training

AI Rational's research demonstrated the effectiveness of transformer models, specifically highlighting the superior performance of RoBERTa over BERT and DistilBERT in their setups [37]. Motivated by their findings, we chose RoBERTa due to its demonstrated efficacy and additionally trained five other models of our choice: SocBERT[38], Covid-Twitter-BERT[39], BioBERT-v1.1[40], ClinicalBERT[23], and PubmedBERT[41]. Our models were trained with a learning rate of 5×10^{-5}, a batch size of 16, gradient accumulation steps of 2, and 20 epochs. Early stopping with a patience of 2 was used to prevent overfitting.

6.1.5 Results and Discussion

This section presents detailed results for the six models trained to address the three tasks of misinformation detection, focusing on their F1 scores and overall accuracy.

6.1.5.1 Task 1A: Check-Worthiness of Tweets

The table below presents the F1 scores and accuracy for each model evaluated in Task 1A. The results are compared against the baseline performance established by AI Rational, the top performer in the CLEF-2022 challenge, which achieved an F1 score of 0.70 and an accuracy of 0.84. In this task, F1 score of the positive class was used for ranking.

Model	**F1 Score (Check-Worthy)**	**Accuracy**
SocBERT	0.51	0.79
Covid-Twitter-BERT	0.71	0.87
BioBERT-v1.1	0.49	0.80
ClinicalBERT	0.50	0.78
PubmedBERT	0.57	0.83
RoBERTa	0.56	0.81
Baseline (AI Rational)	**0.70**	**0.84**

TABLE 6.2: F1 scores and accuracy of models for Task 1A

Discussion of Results: The results indicate that Covid-Twitter-BERT outperforms the baseline with an F1 score of 0.71 with respect to the minority class and an accuracy of 0.87, suggesting that this model is particularly effective at identifying tweets that should be checked for veracity. This performance could be attributed to its training on a dataset closely related to Twitter communications during the COVID-19 pandemic, which may have provided it with a nuanced understanding of the types of misinformation prevalent on social media.

Other models, such as BioBERT-v1.1, ClinicalBERT, and RoBERTa, showed lower performance compared to the baseline. These models, although powerful in biomedical and general contexts, may not be as effective for the specific nuances of Twitter data, which includes slang, abbreviations, and a dynamic use of language.

SocBERT and PubmedBERT also performed under the baseline but showed better adaptability compared to BioBERT-v1.1 and ClinicalBERT. These results highlight the challenges of domain adaptation and the importance of model training that aligns closely with the target application's data characteristics.

6.1.5.2 Task 1B: Verifiable Factual Claims Detection

The results for Task 1B are summarized in the table below, which presents the F1 scores and accuracy for each model evaluated. This task focused on detecting verifiable factual claims, with RoBERTa showing the highest performance. The baseline in this context is the performance of PoliMi-FlatEarthers, which utilized a system based on GPT-3.

Model	**F1 Score (Not Verifiable)**	**Accuracy**
SocBERT	0.72	0.80
Covid-Twitter-BERT	0.73	0.79
BioBERT-v1.1	0.71	0.78
ClinicalBERT	0.72	0.77
PubmedBERT	0.71	0.79
RoBERTa	0.76	0.82
Baseline (PoliMi-FlatEarthers)	**-**	**0.76**

TABLE 6.3: F1 scores and accuracy of models for Task 1B

Discussion of Results: RoBERTa outperformed all other models as well as the baseline, with an F1 score of 0.76 and an accuracy of 0.82. This superior performance

can be attributed to RoBERTa's robust optimization techniques and its ability to effectively handle the nuances of diverse datasets, which is crucial for tasks requiring the identification of verifiable content.

Other models, including SocBERT, Covid-Twitter-BERT, and ClinicalBERT, also performed above the baseline, showcasing their respective strengths in dealing with factual content. However, they slightly trailed behind RoBERTa, underscoring the latter's suitability for complex NLP tasks that involve subtleties and varied contexts.

The baseline set by PoliMi-FlatEarthers, although utilizing the advanced GPT-3 model, achieved an accuracy of 0.76. This highlights the challenges even advanced models like GPT-3 face in specific NLP tasks like verifiable claim detection. It also emphasizes the need for specialized training or tuning to enhance performance in specialized tasks.

6.1.5.3 Task 1C: Harmful Tweet Detection

This task focused on detecting harmful tweets, a challenge compounded by significant class imbalance. Below is a summary of the F1 scores and accuracy for each model, compared against the baseline established by Team Zorros, which employed an ensemble of five transformer models.

Model	**F1 Score (Harmful)**	**Accuracy**
SocBERT	0.15	0.91
Covid-Twitter-BERT	0.19	0.90
BioBERT-v1.1	0.40	0.89
ClinicalBERT	0.22	0.91
PubmedBERT	0.35	0.89
RoBERTa	0.31	0.90
Baseline (Team Zorros)	**0.397**	**0.68**

Table 6.4: F1 scores and accuracy of models for Task 1C

Discussion of Results: Among the models tested, BioBERT-v1.1 demonstrated the highest F1 score of 0.40, outperforming the baseline's F1 score. This achievement is particularly notable given the severe class imbalance, which often depresses performance metrics such as F1 scores. The accuracy of BioBERT-v1.1, while slightly lower than some other models, suggests a reasonable trade-off between precision and recall.

The baseline performance by Team Zorros, although achieving an F1 score close to that of BioBERT-v1.1, had significantly lower accuracy. This indicates that while their ensemble approach was effective at identifying harmful content, it may have suffered from a higher rate of false positives or negatives compared to single-model approaches.

6.1.5.4 Impact of Data Imbalance

The class imbalance has a pronounced effect on the F1 scores, particularly for the minority class which these scores represent. In cases where the "No" class significantly outweighs the "Yes" class, as in Task 1C, even a model with high accuracy may have a poor F1 score, indicating a large number of false negatives. This issue highlights the challenge of training models on skewed datasets where the presence of critical but rare classes can skew performance metrics.

6.1.5.5 Strategic Implications for Future Work

Future work will need to address these disparities through strategic application of data sampling techniques such as oversampling the minority class or undersampling the majority class to provide a more balanced dataset for training. Additionally, employing cost-sensitive learning that penalizes the misclassification of the minority

class could improve performance. Advanced techniques such as synthetic data generation using approaches like SMOTE might also be explored to enhance the model training process and mitigate the impact of data imbalance.

In conclusion, while the models show promising results in certain areas, their effectiveness is curtailed by the inherent challenges posed by data imbalance. Addressing these issues will be crucial for improving the robustness and reliability of misinformation detection systems in real-world applications.

6.2 Veracity Prediction Pipeline

6.2.1 Methodology

To address the challenge of veracity prediction, we utilized the PubHealth dataset, focusing specifically on claims labeled as "True" and "False". Inspired by the methodologies detailed in the study by Kotonya and Toni (2020) [42], we employed cosine similarity to assess the relevance of evidence to the claims, thereby enhancing the prediction process.

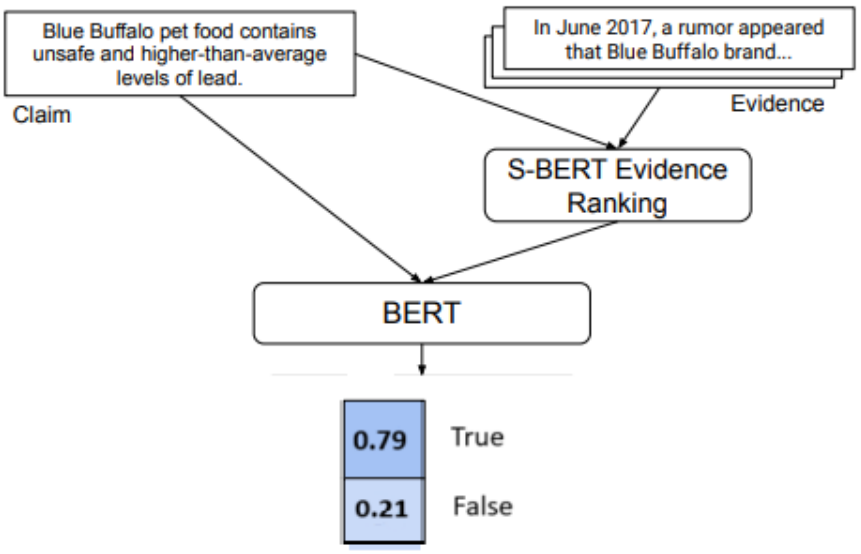


FIGURE 6.2: Architecture of veracity prediction

In our implementation, we first encoded the claims and potential evidence sentences derived from the main-text of the claim using Sentence-BERT (S-BERT), designed to capture sentence-level semantic meanings effectively. By computing the cosine similarity between encoded claims and evidence sentences, we could rank the sentences based on their relevance. The top 5 ranked sentences along with the claim were then used as inputs to our veracity prediction models.

6.2.2 Model Training

Following the approach outlined by Kotonya and Toni, we trained four specific models on the PubHealth dataset, each known for their strengths in processing biomedical text and language understanding:

- **BERT (Bidirectional Encoder Representations from Transformers)**: Developed by Devlin et al., BERT revolutionized the way contextual information is integrated into language models by using a mechanism known as attention, predicting masked tokens in a sequence to learn context.

- **BioBERT v1.0 and v1.1**: These are variants of BERT fine-tuned on biomedical literature, including PubMed abstracts and PubMed Central articles. BioBERT has shown improvements in tasks requiring biomedical domain knowledge.
- **SciBERT**: Another BERT derivative, SciBERT is pre-trained on a large corpus of scientific text, making it suitable for tasks involving scientific terminology and concepts.

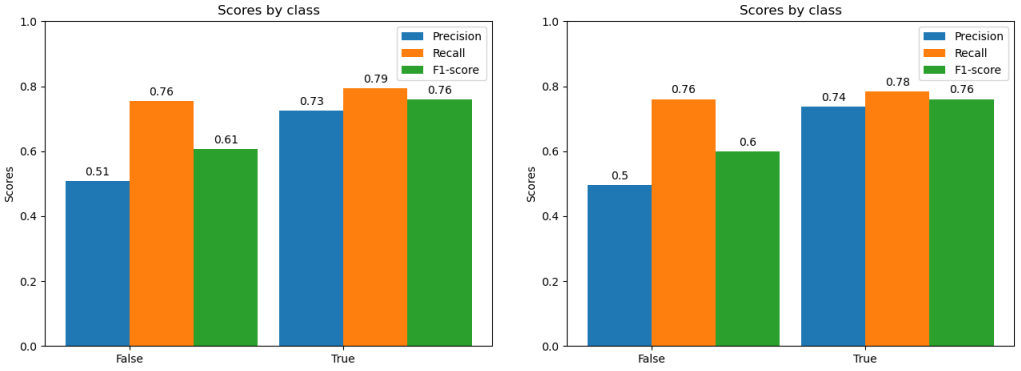
These models were selected due to their proven efficacy in domain-specific contexts, as demonstrated by prior research. Training was focused solely on binary classification tasks — distinguishing between "True" and "False" claims to ensure high precision in public health information dissemination.

6.2.3 Performance Metrics

The performance of each model was assessed on the PubHealth dataset for veracity prediction, evaluating precision, recall, and F1-score for both "True" and "False" claim categories. These metrics are crucial for understanding the effectiveness of each model in distinguishing between verifiable and non-verifiable claims, particularly pertinent in the context of public health misinformation where accurate information is crucial.

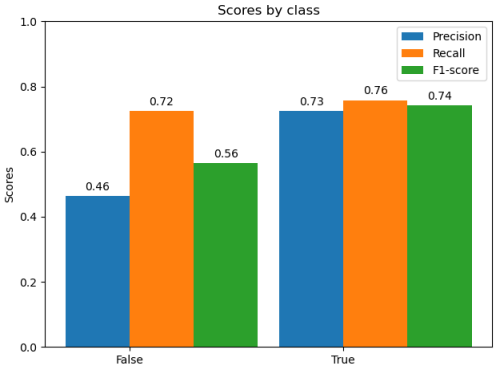
6.2.4 Discussion

The performance metrics presented in Figure 6.3 and detailed above provide significant insights into the efficacy of domain-specific models in addressing the critical task of public health fact-checking. The graphical and tabular data representation clearly delineates the strengths and weaknesses of each model in distinguishing between true and false claims.

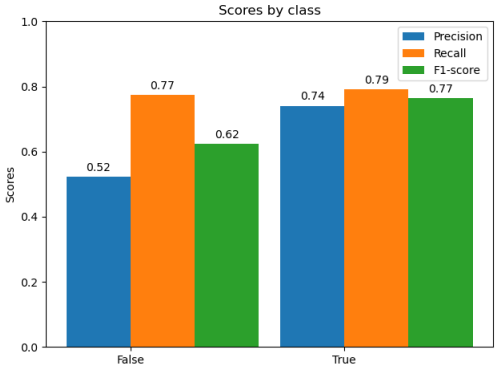


(a) BERT

(b) BioBERT v1.0



(c) BioBERT v1.1



(d) SciBERT

FIGURE 6.3: Detailed performance metrics of the models on the PubHealth dataset for veracity prediction.

Notably, the variations in performance across models suggest that certain architectures may be more adept at processing the nuanced language typical of medical literature and public health information. The higher performance scores for "True" categories across models emphasize the importance of precision in correctly identifying verifiable information, which is paramount in preventing the spread of misinformation in public health contexts.

Chapter 7

Conclusion and Future Works

7.1 Conclusion

This thesis has illuminated the formidable capabilities of Large Language Models (LLMs) in both generating and detecting misinformation. The development and deployment of an enhanced detection pipeline, which now includes sophisticated tasks such as check-worthiness detection, factual verification, and harm detection, marks a significant advancement in our ability to combat misinformation.

These additions have substantially bolstered our detection framework, showcasing the robustness of current defense mechanisms against a complex landscape of misinformation. Through systematic testing, the majority of misinformation instances crafted by LLMs were effectively identified and mitigated, demonstrating the efficacy of the AI community's collective efforts.

However, the research also brought to light the severe consequences that even a single undetected piece of misinformation can have. Instances like the baseless linking of 5G technology with COVID-19, which led to real-world violence and panic, underline the critical need for highly reliable detection systems. Such scenarios vividly

demonstrate the potential havoc that misinformation can wreak, emphasizing the importance of precision in detection technologies.

The intrinsic limitations of the datasets used, particularly the AAAI Constraint dataset, point to an ongoing challenge: prevalent misinformation types are well-represented, whereas rarer or emerging forms are underrepresented. This can potentially lead to gaps in detection coverage. Additionally, our findings underscore the adaptability of LLMs, which can vary their output based on the framing of prompts, suggesting that misinformation generators might employ indirect methods that are harder to detect.

7.2 Future Works

As we pivot towards the avenues yet to be explored in the vast domain of misinformation detection, one emerging concern is the propagation of 'half-truths.' These are craftily devised statements where verifiable truths are woven seamlessly with falsehoods, often presenting a skewed narrative that can mislead the public while eluding detection systems. [43]

Consider this example: "Renowned scientists at XYZ University have proven that consuming vitamin C boosts immunity. Therefore, high doses of vitamin C can act as an alternative to COVID-19 vaccinations." In the aforementioned statement, while it is scientifically accepted that vitamin C plays a role in boosting immunity, suggesting it as an alternative to COVID-19 vaccinations is misleading and potentially harmful. The initial true statement lends credibility to the subsequent misinformation, creating a potent combination that can both deceive the public and confound detection models.

Exploring ways to detect and handle these half-truths should be at the forefront of future endeavors. This would necessitate developing more nuanced and context-aware detection algorithms capable of dissecting such statements, understanding the implications of each segment, and making a holistic decision regarding its veracity.

Another aspect worth delving into is the iterative training of detection models using misinformation generated by advanced language models like LLM. By consistently exposing defense mechanisms to the latest techniques and patterns in misinformation generation, we could foster a continuously evolving detection system, always staying one step ahead of potential deceptive narratives.

Lastly, expanding the scope and diversity of datasets used for training and testing is imperative. While prominent misinformation can be tracked and countered, lesser-known deceptive narratives often remain undetected. Incorporating a wider range of data sources, perhaps even leveraging real-time data streams, could enhance the robustness and adaptability of our defense mechanisms.

Bibliography

[1] J. Introne, Irem Gokce Yildirim, L. Iandoli, J. DeCook, and Shaima Elzeini. How people weave online information into pseudoknowledge. *Social Media + Society*, 4(3):1–12, 2018.

[2] H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236, 2017.

[3] X. Zhou and R. Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *arXiv preprint arXiv:1812.00315*, 2018.

[4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[5] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2931–2937, 2019.

[6] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.

[7] Gautam Kishore Shahi, Anne Dirkson, Nisarg Majumder, Abdul Jalal, and Karin Verspoor. Recovery: A multimodal repository for covid-19 news credibility research. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3469–3472. ACM, 2020.

[8] Chuizheng Meng Sirisha Rambhatla Yan Liu Department of Computer Science University of Southern California Los Angeles USA. Karishma Sharma, Sungyong Seo. Covid-19 on social media: Analyzing misinformation in twitter conversations a preprint. Unknown.

[9] MOHAMED TAHA AHMED TAHA ESSAM H. HOUSSEIN DIAA SALAMA ABDELMINAAM, FATMA HELMY ISMAIL and AYMAN NABIL. Coaid-deep: An optimized intelligent framework for automated detecting covid-19 misleading information on twitter. Unknown.

[10] Kai Shu Huan Liu Yichuan Li, Bohan Jiang. Mm-covid: A multilingual and multimodal data repository for combating covid-19 disinformation. Unknown.

[11] M.A. Serhani b I. Taleb a S.S. Mathew a K. Hayawi a, S. Shahriar a. Anti-vax: a novel twitter dataset for covid-19 vaccine misinformation detection. Unknown.

[12] Xiaofeng Yan Tianqi Yang Wenshuo Wang Zehao Huang Xiongye Xiao Shahin Nazarian Paul Bogdan Mingxi Cheng, Songli Wang. A covid-19 rumor dataset. Unknown.

[13] Yashoda Barve and Jatinderkumar R. Saini. Detecting and classifying online health misinformation with 'content similarity measure (csm)' algorithm: an automated fact-checking-based approach. Unknown.

[14] Gautam Kishore Shahi and Durgesh Nandini. Fakecovid- a multilingual cross-domain fact check news dataset for covid-19. Unknown.

[15] Amirhosein Damia Razieh Bahmanyar Mohammadreza Parvizimosaed, Mehdi Esnaashari. Using supervised learning models for creating a new fake

news analysis and classification of a covid-19 dataset: A case study on covid-19 in iran. Unknown.

[16] Arjuna Ugarte Yoshitomo Matsubara Sean Young Sameer Singh Tamanna Hossain, Robert L. Logan IV. Covidlies: Detecting covid-19 misinformation on social media. Unknown.

[17] H. Miwa (Eds.) L. Barolli, K.F. Li. Covid-19-fakes: a twitter (arabic/english) dataset for detecting misleading information on covid-19. Unknown.

[18] Tuhin Chakrabarty Arkadiy Saakyan1 and Smaranda Muresan. Covid-fact: Fact extraction and verification of real-world claims on covid-19 pandemic. Unknown.

[19] Roman Klinger Isabelle Mohr, Amelie Wührl. Covert (a corpus of fact-checked biomedical covid-19 tweets). Unknown.

[20] Ayan Basak Sourya Dipta Das and Saikat Dutta. A heuristic-driven ensemble framework for covid-19 fake news detection. 2021.

[21] nan. Fakehealth dataset. Unknown.

[22] Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. *arXiv preprint arXiv:2102.07350*, 2021.

[23] Mark Chen et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[24] J. E. Hu et al. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[25] Susan Zhang et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[26] Andy Zou, Zifan Wang, J. Z. Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

[27] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing nlp. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[28] Minhao Cheng, Jinfeng Le, Xinyi Yi, and Ralph Grishman. Imitation attacks and defenses for black-box machine translation systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019*, 2019.

[30] Jens Lehmann et al. Dbpedia: A large-scale, multilingual knowledge base extracted from wikipedia. 2015.

[31] Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. In *Communications of the ACM*, volume 57, pages 78–85. ACM, 2014.

[32] Stephan Lewandowsky, Ullrich K H Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3):106–131, 2012.

[33] Naeemul Hassan, Bill Adair, James T. Hamilton, Chengkai Li, Mark Tremayne, and Jun Yang. The quest to automate fact-checking. *Proceedings of the 2015 Computation+Journalism Symposium*, 2015.

[34] Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. Predicting factuality of reporting and bias of news media sources. *EMNLP 2018*, 2018.

[35] Neema Kotonya and Francesca Toni. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online, November 2020. Association for Computational Linguistics.

[36] Preslav Nakov, Shaden Shaar, Firoj Alam, Giovanni Da San Martino, Lluís Márquez, Alberto Barrón-Cedeño, and Tamer Elsayed. Overview of the clef-2022 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *CLEF 2022 Conference and Labs of the Evaluation Forum*. CLEF, 2022.

[37] AI Rational Team. Ai rational at clef-2022 checkthat!: Employing transformer models for detecting misinformation in tweets. In *CLEF 2022 Conference and Labs of the Evaluation Forum*. CLEF, 2022.

[38] Yuting Guo and A. Sarker. Socbert: A pretrained model for social media text. *ACL Anthology*, 2023.

[39] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, 2020.

[40] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[41] Emily Alsentzer, John Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, 2019.

[42] Neema Kotonya and Francesca Toni. Explainable automated fact-checking: A survey. *Computational Linguistics*, 46(2):353–395, 2020.

[43] Sameep Mehta Varad Bhatnagar Pushpak Bhattacharyya Sandeep Singamsetty, Nishtha Madaan. "beware of deception": Detecting half-truth and debunking it through controlled claim editing. 2023.