Enhancing Propaganda Detection Using Synthetic Data Informed by Human Disinformation Strategies

A Report Submitted in
Partial Fulfilment of the Requirements
for the B.Tech Final Year Project

by

Deep Saikia (Sc ID 2112129)

Aryan Kumar Singh (Sc ID 2112161)

Diwesh Tiwari (Sc ID 2112044)

Md Shohan Mia (Sc ID 2112161)

Under the Supervision of

Dr. Debbrota Paul Chowdhury

Dept. of Computer Science and Engineering
NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR
May, 2025

COMPUTER SCIENCE & ENGINEERING DEPARTMENT
NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR
(An Institute of National Importance)
SILCHAR, ASSAM, INDIA – 788010
Fax: (03842) 224797 **Website:** http://www.nits.ac.in

Declaration

Project Title: **Enhancing Propaganda Detection Using Synthetic Data Informed by Human Disinformation Strategies**
Degree for which the Thesis is submitted: **Bachelor of Technology**

We declare that the presented thesis represents largely my own ideas and work in my own words. Where others ideas or words have been included, We have adequately cited and listed in the reference materials. The thesis has been prepared without resorting to plagiarism. We have adhered to all principles of academic honesty and integrity. No falsified or fabricated data have been presented in the thesis. We understand that any violation of the above will cause for disciplinary action by the Institute, including revoking the conferred degree, if conferred, and can also evoke penal action from the sources which have not been properly cited or from whom proper permission has not been taken.

Date:
Signature:

Deep Saikia 2112129
Aryan Kumar Singh 2112161
Diwesh Tiwari 2112044
Md Shohan Mia 2112161

COMPUTER SCIENCE & ENGINEERING DEPARTMENT
NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR
(An Institute of National Importance)
SILCHAR, ASSAM, INDIA – 788010
Fax: (03842) 224797 **Website:** http://www.nits.ac.in

Certificate

It is certified that the work contained in this project report entitled "**Enhancing Propaganda Detection Using Synthetic Data Informed by Human Disinformation Strategies**" submitted by **Deep Saikia, Aryan Kumar Singh, Diwesh Tiwari and Md Shohan Mia** for the B.Tech. project is absolutely based on their own work carried out under my supervision.

Place:
Date:

Dr. Debbrota Paul Chowdhury
Computer Science & Engineering
National Institute of Technology Silchar

"You have to dream before your dreams can come true."

A. P. J. Abdul Kalam

Abstract

This thesis presents the development and application of a comprehensive framework for detecting propaganda in news articles, combining synthetic data generation with advanced deep learning techniques that integrate contextual and relational modeling.

Propaganda, deliberate attempts to shape public opinion through misleading or deceptive means, has evolved significantly in the digital age, driven by social media and AI-generated disinformation. Traditional methods for detecting propaganda often rely on manually curated datasets or crowdsourced annotations. However, these approaches are labor-intensive, susceptible to bias, and struggle to keep up with the ever-changing tactics used in modern propaganda. Additionally, most existing classifiers analyze articles in isolation, overlooking the connections between different pieces of propagandistic content.

To address these challenges, this paper introduces a synthetic dataset generation pipeline that systematically embeds human-like propaganda techniques into factual news articles. Using a diverse collection of real news stories, we first apply extractive summarization to identify key sentences. These sentences are then modified using a sequence-to-sequence model, incorporating propaganda strategies such as Loaded Language, Appeal to Fear,Flag Waving, etc. The resulting dataset, PropDataSet, serves as training material for a proposed hybrid classification model that combines BERT for deep linguistic analysis and a Graph Attention Network (GAT) to capture relationships between articles.

This research makes two key contributions: the creation of a propaganda-enriched dataset and the development of a hybrid model for detecting disinformation. By integrating automated dataset generation with advanced neural architectures, our approach improves both the adaptability and robustness of propaganda detection models. The combination of BERT's language processing capabilities and GAT's ability to model inter-article relationships enables a more sophisticated understanding of propaganda, surpassing traditional classification techniques. This study underscores the value of synthetic data-driven training in the fight against misinformation, offering a scalable and adaptable solution for detecting disinformation in digital media.

Acknowledgements

We would like to express our deepest gratitude to the following individuals who have contributed significantly to the completion of this project.

First and foremost, we would like to thank our project supervisor, **Dr. Debbrota Paul Chowdhury** Sir, for his guidance, support, and encouragement throughout the project. His expertise and valuable feedback have been instrumental in shaping our ideas and improving our work.

Additionally, we would like to thank the **Department of Computer Science and Engineering of National Institute of Technology Silchar** for providing us with the necessary resources and facilities to complete this project.

Lastly, we would like to thank each other for the collaborative effort and teamwork that has made this project possible.

Date:

Signature:

Deep Saikia

Aryan Kumar Singh

Diwesh Tiwari

Md Shohan Mia

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
BART	Bidirectional and Auto-Regressive Transformers
CNN	Convolutional Neural Network
CRF	Conditional Random Field
GPT	Generative Pre-trained Transformer
LLM	Large Language Model
LSTM	Long Short-Term Memory
NLI	Natural Language Inference
NLP	Natural Language Processing
RoBERTa	Robustly Optimized BERT Approach
SCST	Self-Critical Sequence Training
TF-IDF	Term Frequency-Inverse Document Frequency
BERT	Bidirectional Encoder Representations from Transformers
F1	F1 Score (harmonic mean of precision and recall)
GAT	Graph Attention Network
GCN	Graph Convolutional Network
MLM	Masked Language Modeling

CHAPTER 1

Introduction

1.1 Background

Propaganda refers to deliberate attempts by individuals or groups to influence public opinion or behavior toward a predetermined goal, often through manipulative and deceptive methods [1]. While propaganda has existed in various forms for decades, modern disinformation campaigns have leveraged social media and advanced digital platforms to reach unprecedented scales of influence [2]. The digital age has facilitated the widespread dissemination of false information disguised as legitimate news, which creates a significant challenge for detection. Disinformation often imitates real news, making it difficult to distinguish fact from fiction, especially when propaganda techniques, such as appeals to emotion or logical fallacies, are employed [3].

The rise of social media as a dominant news source has further intensified this issue, offering an unregulated environment in which fake news can easily proliferate [3]. Individuals increasingly consume news through social platforms, often without critically evaluating its authenticity, leading to measurable real-world consequences [4, 5]. Compounding the problem, AI-generated content now enables the mass production of disinformation with increasing realism and scale, enhancing the ability of malicious actors to deceive and manipulate public perception [6]. These developments underscore the urgent need for sophisticated propaganda detection systems that go beyond superficial content analysis and can adapt to evolving disinformation tactics.

Early research efforts in propaganda detection largely focused on the news outlet level [7, 8], wherein entire sources were labeled as propagandistic or not. However, this approach often introduces noise: propagandistic sources sometimes publish neutral content to enhance credibility, while reputable outlets occasionally disseminate biased narratives [9]. As a result, article-level or sentence-level detection is required to achieve a finer-grained and more accurate assessment.

Efforts to obtain article-level or sentence-level annotations through crowdsourcing have met challenges, as annotators' personal beliefs may influence their judgment, particularly when content aligns with their worldview. Consequently, professionally curated datasets, focusing on propaganda techniques at the fragment level, are considered more reliable and precise for training detection models.

Traditional methods of collecting disinformation data,such as scraping articles from low-credibility sources or using fact-checking websites, face limitations in reliability, scalability, and coverage. Articles from unreliable outlets may not always contain disinformation, and misleading content is frequently removed, making such datasets volatile. Fact-checking-based annotations, while valuable, are labor-intensive and limited in scale [3].

To address these constraints, our work proposes a novel methodology for generating high-quality synthetic disinformation datasets. By combining rhetorical propaganda strategies with controlled factual modifications, and leveraging models like BART [10] for sentence rewriting, we generated realistic examples of disinformation augmented with propaganda techniques such as flag-waving, loaded language, appeal to fear, and appeal to authority. This dataset, termed **PropDataSet**, serves as a scalable alternative to manual annotations.

Building upon that foundation, this research introduces a hybrid propaganda detection model that combines deep contextual embeddings with relational learning. Specifically, BERT [11] is employed to capture rich semantic representations of individual text segments, while Graph Attention Networks (GAT) [12] model inter-article relationships through a similarity graph. This architecture enables the model to reason about not only local linguistic cues but also broader contextual patterns across documents. Fine-tuned using the PropDataSet, the hybrid model demonstrates improved classification performance, especially in detecting subtle or coordinated propaganda tactics.

Together, these two contributions, a scalable, technique-informed synthetic dataset and a context-aware hybrid detection model, offer a comprehensive and adaptable framework for addressing the multifaceted challenge of propaganda detection in the digital age.

1.2 Problem Statement

Current Challenges: The detection of propaganda in digital media remains a complex challenge due to several intertwined factors. Traditional models often rely on labeling entire news articles or sources as propagandistic, a method that introduces significant noise and overlooks the nuanced use of persuasive tactics within individual text fragments. Crowdsourced annotation, while scalable, tends to suffer from subjectivity and inconsistency, as annotators may unconsciously project their own biases, especially when political or emotionally charged content is involved. Meanwhile, datasets curated from unreliable news sources are frequently inconsistent or outdated, and those based on manual fact-checking are limited in size and costly to produce. Furthermore, most existing detection models focus heavily on factual correctness, neglecting the rhetorical techniques, such as appeal to authority, loaded language, or flag-waving, that characterize human-crafted propaganda. Finally, even advanced language models like BERT are often limited by their article-level scope, struggling to capture broader semantic or relational patterns that span multiple documents or campaigns.

Our Focus: To address these issues holistically, this thesis is structured in two complementary phases: In the first phase, we designed a synthetic disinformation generation pipeline that constructs a high-quality dataset (**PropDataSet**) by embedding controlled factual inaccuracies and rhetorical propaganda strategies into authentic news articles. Using models like BART for sequence-to-sequence generation and incorporating self-critical sequence training (SCST) with entailment checks, this dataset aims to simulate real-world propaganda in a scalable, flexible, and ethically controlled manner. By mimicking the structure and tone of human-authored fake news, the dataset serves as a robust training resource for downstream classification tasks.

In the second phase, we developed a hybrid detection framework that leverages the strengths of two complementary models: BERT, for deep linguistic and contextual

understanding of individual articles, and Graph Attention Networks (GAT), for modeling inter-article semantic relationships. This architecture enables the model to simultaneously capture both local text features and global structural patterns, thereby improving the identification of isolated propaganda techniques as well as coordinated disinformation campaigns. Fine-tuned on the synthetic dataset and evaluated on real-world benchmarks, the model demonstrates significant improvements in classification accuracy and generalizability.

Problem Definition: To design and develop a hybrid propaganda detection system that combines BERT for deep semantic understanding and Graph Attention Networks for relational modeling, trained on a synthetically generated dataset crafted using human disinformation strategies to enhance classification accuracy at the article level.

1.3 Objectives

The primary objectives of this thesis are outlined below:

- **Design a synthetic data generation pipeline:** Develop a controlled, scalable approach for generating disinformation-rich content by leveraging advanced NLP models (e.g., BART), capable of modifying key textual segments while embedding propaganda techniques such as appeal to authority, loaded language, and flag-waving.
- **Preserve coherence while ensuring semantic divergence:** Introduce a refinement mechanism using self-critical sequence training (SCST) and natural language inference (NLI) to ensure that generated content is semantically distinct yet contextually coherent with the original text.
- **Construct a high-quality synthetic propaganda dataset:** Collect authentic news articles from reliable sources and process them through the generation pipeline to create a balanced and realistic dataset (**PropDataSet**) for supervised propaganda classification.

- **Develop a hybrid propaganda detection model:** Build a classification framework that integrates BERT for deep contextual understanding and Graph Attention Networks (GAT) for relational modeling between articles.
- **Fine-tune the detection model using PropDataSet:** Train and optimize the BERT+GAT model using the synthetic dataset, enabling the detection of both isolated propaganda techniques and broader patterns of coordinated disinformation.
- **Evaluate model performance across domains:** Assess the proposed model's performance using both synthetic and real-world datasets (e.g., WELFake, MisInfoSet) through accuracy, precision, recall, F1-score, and cross-domain generalization metrics.

1.4 Organization of the Project Report

The report is structured to present the problem, methodology, and contributions in a clear and sequential manner, covering both the synthetic data generation and the development of the hybrid propaganda detection model. It is organized into six main chapters as described below:

Chapter 1: Introduction
This chapter introduces the problem of propaganda in digital media, outlines the motivation behind the study, defines the problem statement and research objectives, and provides a high-level overview of the proposed approach.

Chapter 2: Literature Review
This chapter reviews existing work on disinformation and propaganda detection, including approaches based on natural language processing, synthetic data generation, and neural classification models. It also identifies current limitations and research gaps that this project aims to address.

Chapter 3: Methodology – Part I: Disinformation Dataset Generation
This chapter presents the methodology used to construct the synthetic propaganda

dataset (PropDataSet). It covers article collection, sentence extraction, content modification using sequence-to-sequence models, integration of propaganda techniques, and dataset validation strategies.

Chapter 4: Methodology – Part II: Propaganda Detection Model
This chapter details the architecture of the proposed hybrid detection model. It explains the integration of BERT for textual understanding and Graph Attention Networks (GAT) for relational learning, along with data preprocessing, graph construction, and training procedures.

Chapter 5: Experimental Results and Discussion
This chapter reports the experimental setup, evaluation metrics, and model performance across multiple datasets. It includes comparative analysis with baseline models and discusses the generalization capabilities and practical implications of the proposed system.

Chapter 6: Conclusion and Future Work
This chapter summarizes the key findings and contributions of the project, reflects on its limitations, and proposes directions for future work, including enhancements to the generation pipeline, multi-modal detection, and deployment considerations.

CHAPTER 2

Literature Review

2.1 Overview of Propaganda Detection Research

The detection and analysis of propaganda, particularly in the digital age, is a growing area of research, driven by the proliferation of misinformation and its impact on public opinion. This review synthesizes the contributions of several key studies, offering insights into their objectives, methodologies, and findings while highlighting the challenges and gaps that remain.

Hannah Rashkin et al. (2017) [13] explored linguistic distinctions between fake and real news, focusing on political fact-checking. Their analysis revealed that fake news is often characterized by subjective and dramatic language, whereas real news employs concrete and assertive expressions. Predictive models trained on PolitiFact's truthfulness ratings showed promise, but struggled with out-of-domain accuracy. These findings underscored the potential of lexical markers for identifying deception while also highlighting the limitations of context-agnostic approaches.

Shehel Yoosuf and Yin "David" Yang (2019) [14] advanced the field by leveraging fine-tuned BERT for token-level classification of 18 propaganda techniques. Although the approach improved detection accuracy, the challenges of class imbalance and overfitting limited its generalizability. Similarly, Pankaj Gupta et al. (2019) [15] developed

neural architectures incorporating CNN, LSTM-CRF, and BERT models. Their ensemble techniques enhanced recall and precision but also highlighted the complexity of optimal feature configurations and the risk of overfitting with limited data. These studies marked significant progress in leveraging neural models for nuanced detection but underscored the importance of robust datasets and careful tuning.

Jinfen Li et al. (2019) [16] demonstrated the potential of Logistic Regression for propaganda detection using linguistic and emotional features alongside TF-IDF and BERT embeddings. Their model achieved an F1 score of 66.16%, illustrating the efficacy of simpler approaches when feature selection is carefully curated. However, their small dataset and sentence-level analysis limited the model's applicability to broader contexts.

On the application front, Alberto Barrón-Cedeño et al. (2019) [7] introduced Proppy, a real-time system designed to detect propaganda in online news. By combining clustering, deduplication, and machine learning, Proppy achieved a remarkable F1 score of 96.72%. However, its reliance on English-only articles and its restriction to a push model revealed gaps in adaptability and user engagement.

The seminal work by Giovanni Da San Martino et al. (2019) [17] introduced a multi-granularity neural network for detecting fine-grained propaganda techniques in news articles. This method outperformed traditional BERT-based models in identifying propaganda fragments, demonstrating the potential of multi-level feature aggregation. However, limitations in annotator agreement and sentence-level classification accuracy underscore the difficulty of manual annotation and the complexity of fragment-level propaganda.

Building on this, the authors later provided a comprehensive survey on computational propaganda detection methods [18]. They explored both text-based and network analysis approaches, highlighting the benefits of combining these methods to detect coordinated disinformation. However, the lack of large-scale datasets and the evolving nature of malicious behavior remain critical challenges for this field.

Rowan Zellers et al. (2019) [19] addressed the emerging threat of AI-generated disinformation through the Grover model. By both generating and detecting disinformation, Grover demonstrated 92% accuracy in identifying its own outputs. Despite its success,

the focus on text-only propaganda limits its application in more complex multimedia disinformation. In a related effort, Seunghak Yu et al. (2021) [20] emphasized interpretability in propaganda detection by combining syntactic, semantic, and pre-trained language model features. This approach improved classification accuracy and provided greater transparency in the decision-making process, aiding media literacy. Yet, the model struggled with rare propaganda techniques and lacked provisions for misleading fact-based disinformation.

Kai Shu et al. (2021) [21] developed FACTGEN, a framework focused on improving factual consistency in synthetic news generation. By incorporating a fact retriever and claim reconstructor, FACTGEN ensured more reliable news synthesis. While this approach advanced factual alignment, its exclusion of propaganda techniques limited its scope for combating manipulative narratives.Vorakit Vorakitphan et al. (2022) [22] proposed the PROTECT pipeline, a comprehensive system for identifying propaganda techniques using a two-step classification approach with RoBERTa. Despite achieving an end-to-end classification pipeline, the model faced misclassification issues due to tokenization challenges and lacked a detailed evaluation of its outputs. Prashanth Vijayaraghavan et al. (2022) [23] focused on propaganda detection in social media with their TWEETSPIN corpus and the MV-PROP model. By leveraging multi-view representations, including semantic, relational, and knowledge features, MV-PROP achieved superior performance compared to baseline models. However, the informal and fragmented nature of Twitter content, coupled with weak annotations, posed challenges for accurate detection and highlighted the need for better data labeling.

The potential of large language models (LLMs) like GPT-3 and GPT-4 was explored by Kilian Sprenkamp et al. (2023) [24]. Their study showed that using "chain of thought" prompting with GPT-4 could match state-of-the-art performance, demonstrating the versatility of LLMs in understanding complex propaganda techniques. However, limitations such as overfitting, hallucinations, and reasoning failures underscore the need for further refinement.

Kung-Hsiang Huang et al. (2023) [25] developed the PROPANEWS dataset to generate realistic propaganda-laden disinformation. By incorporating techniques like loaded language and appeals to authority, the generated articles outperformed baseline models

in realism and detection performance. Nonetheless, the restriction to only two propaganda techniques and the absence of dynamic knowledge integration limit its broader applicability.

2.2 Foundations of the Proposed Hybrid Detection Framework

Recent advancements in deep learning have introduced powerful architectures for both contextual language understanding and relational modeling—capabilities essential for tackling complex disinformation patterns in propaganda detection.

Devlin et al. (2019) [11] introduced BERT (Bidirectional Encoder Representations from Transformers), a transformative model that leverages deep bidirectional context through masked language modeling and next sentence prediction. BERT significantly advanced performance across a wide range of NLP tasks, including text classification benchmarks such as GLUE and SQuAD. Its ability to encode nuanced semantic information made it particularly effective for detecting manipulative language and rhetorical cues in text.

Building on this, Chi et al. (2019) [26] explored best practices for fine-tuning BERT in classification tasks. Their study revealed that task-specific pre-training and optimization strategies—such as multi-task learning and dynamic sampling—could further enhance accuracy, especially in domain-specific applications like propaganda detection.

While BERT excels at understanding individual documents, it lacks the capacity to model inter-document relationships critical for identifying coordinated disinformation. This gap is addressed by the introduction of Graph Neural Networks (GNNs), as proposed by Scarselli et al. (2009) [27], which provide a framework for learning over graph-structured data. GNNs offer powerful mechanisms for capturing both local and global dependencies in a graph, making them well-suited for tasks where relational structure is essential.

Expanding on this, Veličković et al. (2018) [12] proposed Graph Attention Networks (GATs), which incorporate self-attention mechanisms into graph learning. By dynamically weighting the importance of neighboring nodes, GATs enable fine-grained control over information propagation in heterogeneous graphs. This makes them particularly effective for tasks like misinformation detection, where relationships between news articles can indicate underlying patterns of coordinated propaganda.

Together, these foundational works inspired the hybrid architecture developed in this thesis, which combines BERT's deep semantic understanding with GAT's relational modeling to enhance the detection of both standalone and networked propaganda. This integration addresses the limitations of isolated document analysis and facilitates a more holistic view of how disinformation spreads and operates in digital ecosystems.

TABLE 2.1: Overview of key studies on propaganda and fake news detection.

Title	**Objective**	**Methodology**	**Key Findings**
Fine-Grained Analysis of Propaganda in News Articles [17]	Detect and classify specific propaganda techniques in news articles at the fragment level.	Multi-granularity neural network on a manually annotated dataset of 18 propaganda techniques.	Achieved high precision in detecting propaganda fragments but struggled with sentence-level accuracy.
A Survey on Computational Propaganda Detection [18]	Review methods for computational propaganda detection using text and network analysis.	Combines supervised learning with network analysis for malicious behavior detection.	Combining approaches improves detection but lacks large, generalizable datasets.
Defending Against Neural Fake News [19]	Address neural fake news generation with Grover, which also detects AI-generated disinformation.	Trained Grover on news corpora to generate and detect realistic fake news.	Achieved 92% accuracy in detecting AI-generated fake news but limited to text-based analysis.
Interpretable Propaganda Detection [20]	Detect propaganda using interpretable features like syntactic and semantic analysis.	Multi-class classification with semantic and syntactic parsers.	Improved interpretability in propaganda detection but struggled with rare techniques.

TABLE 2.1: Overview of key studies on propaganda and fake news detection.

Title	**Objective**	**Methodology**	**Key Findings**
Fact-Enhanced Synthetic News Generation [21]	Generate fact-enriched synthetic news to ensure factual consistency.	Self-attentive language model with a fact retriever and claim reconstructor.	Enhanced factual consistency but lacked integration of propaganda techniques.
PROTECT: Propaganda Detection [22]	Build a pipeline for automated propaganda detection.	Token-level classification followed by a 14-technique classifier using RoBERTa.	Effective pipeline but misclassified due to tokenization issues.
TWEETSPIN: Fine-Grained Propaganda Detection [23]	Develop a dataset and model for fine-grained propaganda detection on social media.	MV-PROP model leveraging multi-view representations.	Boosted detection accuracy on tweets but suffered from weak annotations.
Large Language Models for Propaganda Detection [24]	Test GPT-3 and GPT-4 for propaganda detection using SemEval-2020 data.	Fine-tuned GPT models with 'chain of thought' prompting.	Comparable to state-of-the-art methods but prone to hallucinations and overfitting.
Faking Fake News for Real Detection [25]	Generate realistic propaganda-based disinformation for fake news detection.	PROPANEWS dataset with self-critical sequence training (SCST).	Created human-like fake news, outperforming baseline models.
Truth of Varying Shades [13]	Analyze language differences in fake and real news.	Lexical analysis using PolitiFact's truthfulness ratings.	Revealed linguistic patterns in fake news but lacked cross-language analysis.
Fine-Grained Propaganda Detection with BERT [14]	Use fine-tuned BERT to classify propaganda techniques.	Token-level classification with oversampling and attention analysis.	Improved detection accuracy but faced overfitting.
Neural Architectures for Propaganda Detection [15]	Develop neural models for propaganda detection.	Combined CNN, LSTM-CRF, and BERT with ensemble methods.	Enhanced precision and recall but risked overfitting due to limited data.
Detection of Propaganda Using Logistic Regression [16]	Classify propagandistic sentences using logistic regression.	Features like TF-IDF, readability, and BERT vectors.	Achieved 66.16% F1 score but struggled with broader context.

TABLE 2.1: Overview of key studies on propaganda and fake news detection.

Title	**Objective**	**Methodology**	**Key Findings**
FakeNewsNet: Data Repository [28]	Create a repository with news, social, and temporal data for fake news research.	Integrated PolitiFact and GossipCop data with social context.	Highlighted diffusion patterns but confined to specific news types.
Proppy: Propaganda Detection System [7]	Real-time propaganda detection system.	Machine learning classifiers with clustering and deduplication.	Achieved 96.72% F1 score but limited to English articles.
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[11]	Introduce BERT, a bidirectional Transformer model, for language representation learning.	Pre-trained BERT using masked language modeling and next sentence prediction, followed by fine-tuning for downstream tasks.	BERT achieved state-of-the-art results on 11 NLP tasks, including GLUE, SQuAD, and SWAG benchmarks.
How to Fine-Tune BERT for Text Classification? [26]	Optimize BERT fine-tuning for text classification.	Tested fine-tuning and multi-task learning on eight datasets.	Improved accuracy with in-domain pre-training.
The Graph Neural Network Model[27]	Propose a Graph Neural Network (GNN) model to process graph-structured data directly.	Introduces a supervised learning algorithm for GNNs and tests on tasks like subgraph matching and mutagenesis.	GNNs effectively handle both node-focused and graph-focused tasks in complex graph-based problems.
Graph Attention Networks [12]	Introduce Graph Attention Networks (GATs) to process graph-structured data using masked self-attention layers.	Uses self-attention mechanisms to assign different weights to node neighbors, with multi-head attention for stability, tested on node classification tasks.	GATs achieved or matched state-of-the-art results on transductive (Cora, Citeseer, Pubmed) and inductive (PPI) benchmarks.

CHAPTER 3

Methodology – Part I: Disinformation Dataset Generation

This chapter presents the methodology used to construct a synthetic propaganda-rich dataset designed to train and evaluate propaganda detection models. Given the limitations of existing datasets, such as annotation bias, limited scale, and lack of rhetorical diversity, this work introduces a controlled generation pipeline that simulates human-authored disinformation. The approach systematically modifies real news articles to embed factual distortions and targeted propaganda techniques while preserving coherence and contextual plausibility.

3.1 News Article Collection

To build a robust and reliable dataset for our project, we began by collecting real news articles from trusted Indian news websites. Our primary goal was to ensure that the dataset represented high-quality journalism and covered topics relevant to the general public. The process of collecting 1,500 articles was conducted in a systematic manner, as described below.

3.1.1 Criteria for Source Selection

When choosing the sources for this dataset, we considered two key criteria. First, the news articles had to originate from credible and trustworthy websites. This was essential to ensure that the content, aside from any intentional manipulations introduced for research purposes, remained authentic. Second, the articles needed to cover events of significant public interest, such as politics, economics, health, or technology, ensuring relevance and diversity in the dataset.

Based on these considerations, we selected well-known Indian news platforms, including *The Hindu*, *The Indian Express*, and *The Times of India*. These outlets have a strong reputation for reliability and journalistic integrity.

3.1.2 Automated Data Collection

The process of collecting articles was automated using a web crawler built with Python, leveraging libraries such as `Scrapy` and `BeautifulSoup`. This crawler systematically navigated the selected news websites to identify and retrieve relevant articles. The key steps in this process included:

- **Identifying Target Pages:** The crawler began by exploring the homepage or sitemap of each website to locate articles published within the last year. This ensured the timeliness of the dataset.
- **Extracting Article Content:** For each identified article, the crawler extracted the title, publication date, main content and other relevant metadata.
- **Filtering Irrelevant Data:** Articles lacking substantial content, such as short updates or promotional material, were excluded to maintain dataset quality.

3.1.3 Data Cleaning and Preprocessing

After collection, the raw data underwent a preprocessing phase to prepare it for use in our study. This phase included:

- **Removing Extraneous Content:** Unnecessary elements such as advertisements, navigation links, and unrelated HTML tags were stripped out.
- **Standardizing Text:** All text was standardized to UTF-8 encoding to handle the diverse language scripts often found in Indian news articles.
- **Language Filtering:** Only articles written in English were retained, as our research focused on generating synthetic data in this language.

3.1.4 Validation and Quality Assurance

To ensure the integrity of the dataset, we conducted a manual review of a randomly selected subset of 100 articles. This review confirmed the accuracy of extracted metadata, the absence of irrelevant content, and the overall quality of the articles. Through this meticulous process, we curated a collection of 1,500 diverse and reliable news articles.

The resulting dataset serves as a strong foundation for our subsequent tasks, particularly in generating synthetic data informed by human disinformation strategies. The careful selection and preparation of this data ensure that our research outcomes are both credible and impactful.

3.2 Article Category Classification

To ensure that the disinformation strategies applied are contextually relevant, the system begins by categorizing each input article into a specific thematic category. This classification step is critical for aligning manipulative techniques with the subject matter of the article, thereby enhancing the credibility and impact of the generated content.

3.2.1 Methodology

We employed a zero-shot classification approach, which is highly versatile as it allows the system to classify text without requiring task-specific training. For this purpose, we used the **facebook/bart-large-mnli** model, a widely recognized transformer-based

model known for its robustness in natural language understanding tasks. The model was accessed via the Hugging Face `pipeline` framework, enabling seamless integration into our system.

A set of predefined categories was defined to cover a broad range of topics. These included themes such as *Science*, *Politics*, *Health*, *Technology*, *Environment*, and others. The system analyzed the textual content of each article and evaluated it against these categories. Based on the model's predictions, the category with the highest confidence score was assigned to the article.

3.2.2 Purpose

The primary objective of this classification step is to ensure that the disinformation strategies employed are contextually appropriate. By aligning the manipulation techniques with the article's theme, we can enhance the believability of the altered content. For instance, an article discussing advancements in renewable energy would be classified under *Environment*, prompting the use of environment-specific disinformation techniques. This thematic alignment ensures that the manipulations resonate with the article's context, increasing their effectiveness.

3.3 Disinformation Generation

3.3.1 Salient Sentence Extraction

After categorizing an article, the next step in our disinformation generation pipeline is the identification of a salient sentence. This sentence serves as the focal point for introducing disinformation. By targeting the most critical or impactful sentence in the article, the system ensures that the manipulation has a substantial effect on the narrative, thereby maximizing its psychological and informational impact.

A salient sentence is typically one that is integral to the overall meaning and coherence of the article. Manipulating or replacing such a sentence can drastically alter the interpretation of the events or arguments presented. By focusing on these pivotal

sentences, we aim to introduce disinformation that significantly disrupts the article's core message while maintaining an appearance of plausibility.

To identify salient sentences, we leveraged extractive summarization techniques. Extractive summarization models are designed to condense an article into its most critical components by selecting sentences that best represent the text's main ideas. For this task, we utilized the `facebook/bart-large-cnn` model, a state-of-the-art summarization framework, to compute the saliency scores of each sentence in an article. These scores indicate the likelihood of a sentence being included in a concise summary of the text.

In practice, the summarization model evaluates the importance of each sentence within the article. The sentence with the highest saliency score, as determined by the model, is identified as the most critical to the article's core message. This sentence is subsequently selected as the target for replacement with generated disinformation.

Identifying salient sentences in news articles poses unique challenges due to the lack of publicly available datasets specifically curated for this task. To address this, we relied on the inherent capabilities of extractive summarization models to estimate saliency. This approach is motivated by the observation that sentences included in summaries are often highly significant. While this method does not explicitly define "saliency," it provides a practical approximation that performs effectively in the context of our study.

For very short articles, where summarization may not yield meaningful results, the entire article is treated as salient. Additionally, fallback mechanisms are in place to handle scenarios where the model encounters unusually large or problematic inputs.

3.3.2 Modifying Salient Sentences through Mask Infilling

After identifying the salient sentence, the next step in the disinformation generation process involves modifying this sentence to subtly embed fabricated elements. This modification shifts the narrative of the article while ensuring that the new content

remains coherent and plausible within the original context. For this purpose, we employed the `facebook/bart-large` model, a fine-tuned version of BART [29], a pre-trained encoder-decoder architecture optimized for tasks like text generation and infilling which was fine tune on loaded langauge,Appeal to fear and flag wavering data from the TWEETSPIN [23] .

The primary goal of this step is to modify the key sentence in a way that introduces disinformation while maintaining the article's coherence. By targeting the most critical sentence in the narrative, we ensure that the manipulation has a substantial impact on the reader's interpretation. The modifications are carefully designed to blend seamlessly with the article, introducing elements such as urgency, exaggeration, or false claims to subtly alter the narrative.

For example, a factual statement such as, *"Deforestation is a growing concern,"* could be transformed into a misleading claim like, *"New reports suggest that deforestation may have already reached irreversible levels."* This transformation introduces a sense of immediacy and alarm, subtly influencing the reader's perception of the issue.

The modification process employs a mask-infilling technique powered by the `facebook/bart-large` model. This involves masking the identified salient sentence within the article's context and generating a rephrased version that introduces the desired disinformative elements. Specifically:

- The original article is first altered by replacing the salient sentence with a placeholder mask. This masked version serves as input to the BART model.
- The model's bidirectional encoder processes the masked text, creating contextualized representations of the article while taking into account the surrounding text.
- Using these representations, the autoregressive decoder generates a new version of the masked sentence. The generation process is guided by the context provided by the rest of the article, ensuring that the output fits seamlessly within the original narrative.

The model's configuration allows it to generate text that aligns with the intended disinformation-propagandic strategy, incorporating subtle elements of bias or fabrication while maintaining a high degree of linguistic coherence.

This step requires a careful balance between introducing disinformation and preserving the article's overall readability and plausibility. BART's advanced architecture, fine-tuned for our specific use case, enables this balance by leveraging its encoder-decoder framework to generate high-quality text that aligns with the context of the article. For articles with particularly short or problematic content, fallback strategies are employed to ensure the modifications remain effective without compromising coherence.

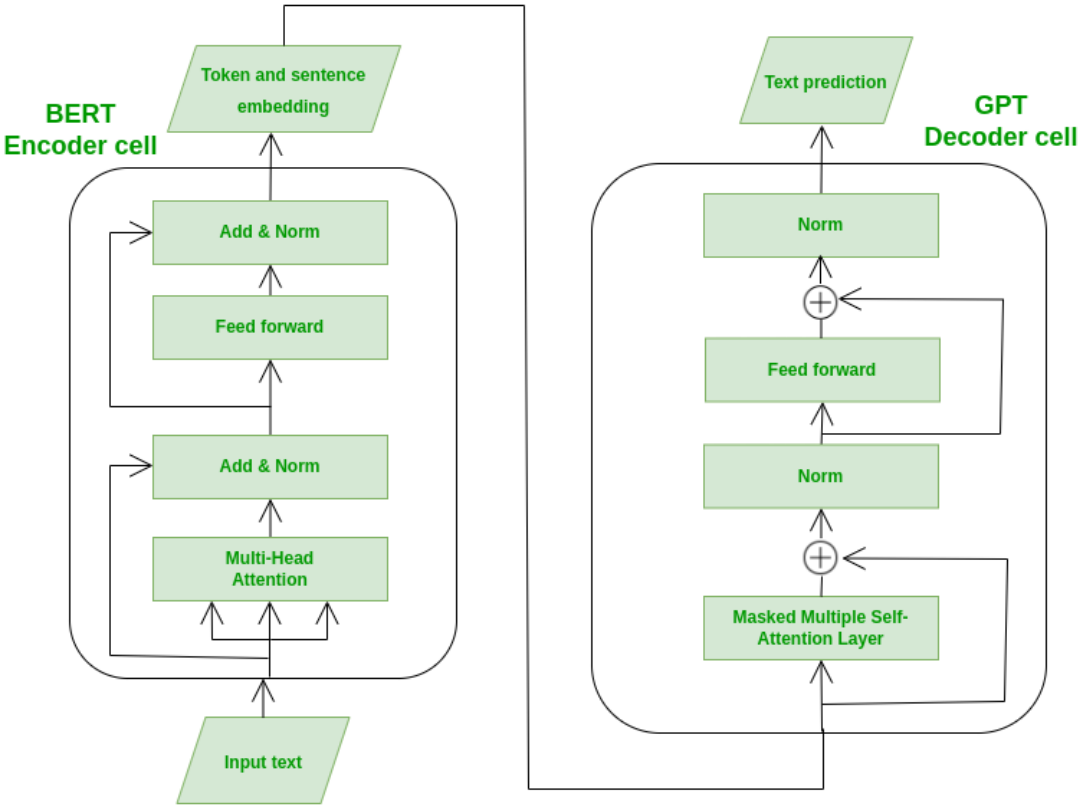

FIGURE 3.1: BART single encoder-decoder network architecture

3.4 Propaganda Techniques Integration

The integration process involves dynamically generating propaganda phrases using a fine-tuned version of the `facebook/bart-large` model and fallback mechanisms based on predefined templates. This ensures that the propaganda seamlessly integrates into the article's context, enhancing its psychological and emotional impact.

3.4.1 Appeal to Authority

The appeal to authority technique is a powerful method for enhancing the credibility of disinformation. By associating misleading narratives with authoritative figures—real, fabricated, or generalized—the system manipulates the reader's inherent trust in expert opinions. This approach exploits cognitive biases, where individuals are more likely to accept information if it is linked to a perceived authority.

To implement this technique, the system leverages Wikidata[1], a structured database of knowledge, to identify authoritative entities relevant to the article's subject matter. It specifically searches for individuals with occupations such as economists, scientists, or public health experts, depending on the thematic category of the article.

To ensure recency and relevance, outdated or less impactful entities are filtered out. For example, the system excludes experts born before 1940 unless their contributions remain highly significant. In scenarios where no specific authorities can be found, placeholders like "renowned expert" or "leading figures in the field" are used to maintain the appearance of credibility.

Once a ranked list of candidates is established, the system generates statements attributed to these entities using a fine-tuned `facebook/bart-large` model. The BART model is trained to seamlessly incorporate authoritative quotes into the text. For example, it can produce outputs like:

> Famous scientist Dr. Jane Doe said "This is the defining challenge of our generation."

[1] https://query.wikidata.org/

To diversify the outputs, multiple templates are employed during the generation process, ensuring that the statements remain contextually appropriate while aligning with the narrative's intended tone.

The appeal to authority serves as a psychological tool to enhance the persuasive power of disinformation. By anchoring fabricated or exaggerated claims to authoritative figures, the system taps into the inherent trust people place in expert opinions. This technique not only increases the perceived legitimacy of the narrative but also reduces the likelihood of skepticism among readers.

Furthermore, the integration of authoritative voices gives the altered content an air of professionalism and sophistication, making it harder for the average reader to identify the information as misleading. This subtle yet effective manipulation reinforces the disinformation's reach and impact, ensuring its acceptance by a broader audience.

3.4.2 Appeal to Fear

Fear-based appeals evoke strong emotional responses by highlighting potential dangers or catastrophic outcomes. If language designed to instill fear and urgency is integrated into the article, it can influence readers' perceptions and decisions.Fear is a powerful motivator, often bypassing rational thought and driving readers toward the intended conclusion.A statement such as, *"Failure to act will lead to catastrophic consequences affecting millions,"* might provoke a sense of urgency and alarm.

3.4.3 Loaded Language

If emotionally charged phrases are employed to subtly bias the reader and evoke strong reactions, these phrases engage readers on an emotional level, often bypassing critical thinking and fostering alignment with the intended narrative.By using emotionally loaded language, the article becomes more engaging and persuasive, subtly guiding the reader's interpretation. An emotionally charged sentence like, *"The world is teetering on the edge of disaster; only immediate action can save us,"* emphasizes the urgency and gravity of the situation, steering the reader's perspective.

3.4.4 Flag Waving

In propagandist articles, Nationalistic or group-centric appeals are integrated to foster a sense of solidarity and pride. This technique resonates with collective identities, particularly in political or social contexts, and aligns the narrative with a shared sense of purpose. By appealing to collective pride or responsibility, this technique strengthens group cohesion and amplifies the message's impact.A phrase such as, *"Our nation must lead the charge to secure a brighter future for all,"* ties the narrative to a sense of national pride and collective duty.

To generate and integrate these three propaganda techniques, the `facebook/bart-large` model was fine-tuned using data from the TWEETSPIN [23] dataset. This dataset contains tweets weakly annotated with 18 fine-grained propaganda techniques. For this work, we selected three specific techniques—Loaded Language, Appeal to Fear, and Flag Waving—and fine-tuned the model accordingly.

The fine-tuned BART model generates disinformation-rich text by prioritizing these styles. During the text generation process:

- The thematic category of the article guides the selection of propaganda styles.
- The model uses context from the article to generate propaganda phrases that blend naturally with the narrative.

The fine-tuned model allows for a sophisticated and dynamic integration of propaganda techniques, enhancing the article's emotional and psychological impact. By combining the generative capabilities of the BART model with the thematic alignment of propaganda techniques, the system ensures a highly effective manipulation of the narrative.

3.5 Reward-Based Refinement

In the mask-infilling step, a reward-based refinement mechanism is employed to ensure that the partially generated content deviates significantly from the original article

while maintaining coherence and readability. This iterative process helps in creating disinformation that is both original and persuasive. The steps involved are as follows:

1. **Initial Generation:** The masked portions of the article are modified using the fine-tuned `facebook/bart-large` model. This serves as the starting point for creating coherent yet misleading content.
2. **Evaluation:** An entailment model evaluates the semantic similarity between the infilled text and the original article. A similarity score is calculated to guide the refinement process.
3. **Reward Calculation:** A reward score is computed based on the similarity score, with penalties for high similarity and rewards for originality.
4. **Iterative Refinement:** Using the reward score as feedback, the system iteratively generates improved versions of the infilled content. The process terminates once an optimal balance between originality and coherence is achieved.

After the mask-infilling step, an additional propaganda technique, i.e. Appeal to Authority, is integrated to enhance the psychological and emotional impact of the disinformation.

3.5.1 Final Reward-Based Refinement

Following the integration of propaganda techniques, a final round of reward-based refinement is applied to ensure the integrated techniques align seamlessly with the overall article. This step focuses on optimizing the final content based on three key metrics:

- **Coherence:** Ensures that the final output is logically consistent across the entire article, including the newly inserted propaganda techniques.
- **Emotional and Psychological Impact:** Verifies that the emotional impact of the disinformation is maximized by refining the psychological triggers, ensuring the manipulative elements resonate with the target audience.

- **Originality and Believability:** Confirms that the article is original enough to be deceptive, yet maintains a sense of believability, so it cannot be easily traced back to the original source.

This final step guarantees that the article is fully refined, ensuring it meets the goals of creating persuasive, emotionally impactful, and original disinformation.

3.5.1.1 Final Reward Calculation Overview

The reward for both intermediate and final refinement is based on three key factors: **Semantic Similarity**, **Emotional Impact**, and **Coherence**. These factors are evaluated separately and combined into a final reward score. Table 3.1 summarizes the key metrics and their significance.

Metric	Purpose	Impact on Reward
Semantic Similarity (S_{sim})	Measures the degree to which the generated article diverges from the original source.	High similarity is penalized to encourage originality.
Emotional Impact (S_{emo})	Evaluates the intensity of emotional and psychological appeals within the text.	Higher emotional scores are rewarded to incentivize persuasive techniques.
Coherence (S_{coh})	Assesses logical consistency, flow, and overall readability of the article.	Higher coherence scores are rewarded to promote well-structured content.

TABLE 3.1: Reward Metrics Used for Article Refinement

3.5.1.2 Calculation of Individual Scores

Each of the metrics is calculated as follows:

1. **Semantic Similarity Score (S_{sim}):** This score is derived from the semantic similarity between the generated article and the original article using a pre-trained language model like `BERT`. A cosine similarity measure is employed to compute the score.

$$S_{sim} = 1 - \text{cosine\_similarity}(\text{embedding}(A_g), \text{embedding}(A_o)) \tag{3.1}$$

Where:

- A_g is the generated article.
- A_o is the original article.

2. **Emotional Impact Score (S_{emo}):** Emotional appeal is quantified by analyzing sentiment and emotional intensity using sentiment analysis models such as `RoBERTa` or a fine-tuned classification model:

$$S_{emo} = \frac{\sum_{i=1}^{n} \text{intensity}(E_i)}{n} \tag{3.2}$$

Where:

- E_i represents emotional categories (e.g., fear, anger).
- $\text{intensity}(E_i)$ is the intensity of the emotional response.

3. **Coherence Score (S_{coh}):** The coherence score is derived using a perplexity measure, calculated by running the article through a language model like `GPT-3`:

$$S_{coh} = \frac{1}{\text{perplexity}(A_g)} \tag{3.3}$$

Where:

- $\text{perplexity}(A_g)$ measures the predictive accuracy of the model on the article.

3.5.1.3 Overall Reward Calculation

The final reward score (R) is a weighted sum of the individual scores:

$$R = \alpha(1 - S_{sim}) + \beta S_{emo} + \gamma S_{coh} \tag{3.4}$$

Where:

- α, β, and γ are the respective weights for semantic similarity, emotional impact, and coherence.
- $1 - S_{sim}$ is used to penalize high similarity, encouraging divergence.

The final score ensures that the content is original, emotionally impactful, and coherent.

3.5.2 Intermediate vs Final Refinement

The difference between intermediate and final refinement lies in the focus of the reward-based process. While the intermediate refinement mainly targets the initial improvement of content coherence, originality, and emotional appeal, the final refinement ensures the alignment of propaganda techniques and guarantees the overall quality of the generated disinformation.

Aspect	**Intermediate Refinement**	**Final Refinement**
Focus	Mask-infilling and semantic variation to introduce meaningful edits.	Integration of propaganda techniques and coherence with emotional appeal.
Coherence	Establishes initial logical consistency and flow within segments.	Ensures full-article coherence and alignment with persuasive intent.
Originality	Encourages deviation from the original article's wording and structure.	Balances originality with plausibility to enhance credibility.
Emotional Impact	Enhances emotional resonance based on topic and context.	Maximizes emotional and psychological triggers through fine-tuned modifications.

Table 3.2: Comparison of Intermediate and Final Refinement Stages

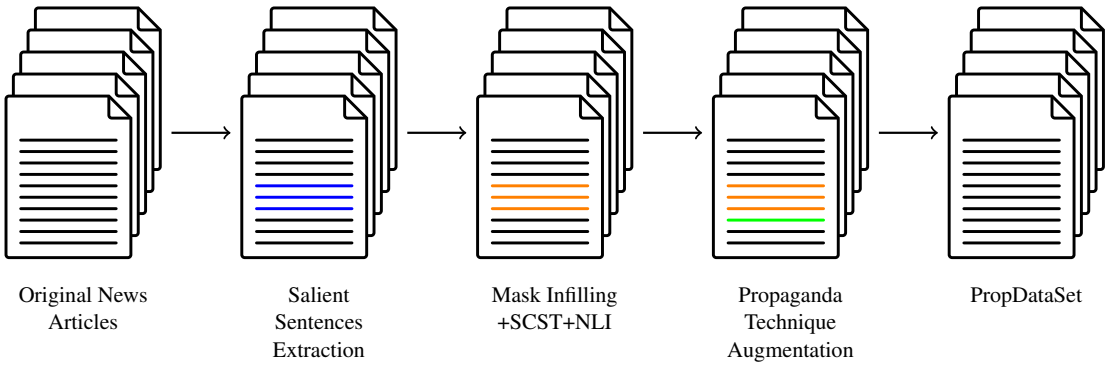

FIGURE 3.2: *Basic overview of generating disinformative news articles.Given a set of authentic news articles, the approach first extracts the* ***Salient sentences****, replaces them with plausible but* ***Misleading sentences*** *using Mask Infilling with SCST and NLI, and finally embeds* ***Propaganda techniques*** *to make the article resemble human-written fake news, resulting in the Propaganda dataset.*

3.5.3 Why a Final Reward-Based Refinement?

The final round of reward-based refinement is crucial because it ensures that all previous modifications—mask-infilling and the integration of propaganda techniques—align cohesively. Without this final refinement, the content may suffer from inconsistencies or an imbalance between emotional impact, coherence, and originality. This step acts as a final quality control layer, guaranteeing that the disinformation achieves maximum psychological influence, emotional resonance, and believability while maintaining the necessary degree of originality. The use of a final refinement step thus ensures that the generated disinformation is both effective and sophisticated, serving its intended purpose without raising suspicion.

3.6 Pipeline Walkthrough: From Input to Output

To demonstrate the disinformation generation pipeline, we provide a step-by-step walkthrough using the following input article as an example:

Scientists have warned that deforestation is a growing concern in many parts of the world. Large areas of forests are being cut down, threatening wildlife and contributing to climate change. Immediate action is required to halt the destruction and preserve biodiversity.

The pipeline processes the input as follows:

1. **Zero-shot Classification:** The article is classified into a thematic category using the `facebook/bart-large-mnli` model. Based on predefined categories such as *Environment*, *Health*, and *Technology*, the model assigns the category with the highest confidence score. **Result:** *Environment.*

2. **Salient Sentence Extraction:** Using `facebook/bart-large-cnn`, an extractive summarization model, the most critical sentence is identified based on saliency scores. **Result:** *Immediate action is required to halt the destruction and preserve biodiversity.*

3. **Mask Infilling:** The salient sentence is replaced with a placeholder mask, and the modified article is processed by the fine-tuned `facebook/bart-large` model. The model generates a disinformative replacement incorporating propaganda techniques. **Result:** *New reports suggest that deforestation may have already reached irreversible levels, and failure to act could result in catastrophic consequences.*

4. **Self-Critical Sequence Training (SCST):** The generated sentence is refined to ensure non-entailment with the original sentence. SCST penalizes similarity and iteratively improves the disinformation content. **Result:** *New reports suggest that deforestation may have already reached irreversible levels, and failure to act could result in catastrophic consequences.*

5. **Propaganda Integration:** The *Appeal to Authority* technique is applied, leveraging fabricated statements from authoritative figures identified using `Wikidata`. **Result:** *Leading environmental scientist Dr. Jane Doe stated, "This is the defining crisis of our time, and failure to act immediately could devastate our planet's future."*

6. **Reward-Based Refinement:** The final refinement step ensures the content is original, coherent, and emotionally impactful by optimizing semantic similarity, coherence, and emotional intensity. **Final Output:** *Scientists have warned that deforestation is a growing concern in many parts of the world. Large areas of forests are being cut down, threatening wildlife and contributing to climate change. New reports suggest that deforestation may have already reached irreversible levels, and failure to act could result in catastrophic consequences. Leading environmental scientist Dr. Jane Doe stated, "This is the defining crisis of our time, and failure to act immediately could devastate our planet's future."*

The pipeline, as shown in figure 3.3, demonstrates how input articles are processed through classification, salient sentence extraction, mask infilling, and propaganda integration to generate a coherent and persuasive disinformative article. Each step is designed to enhance psychological and emotional impact while maintaining plausibility.

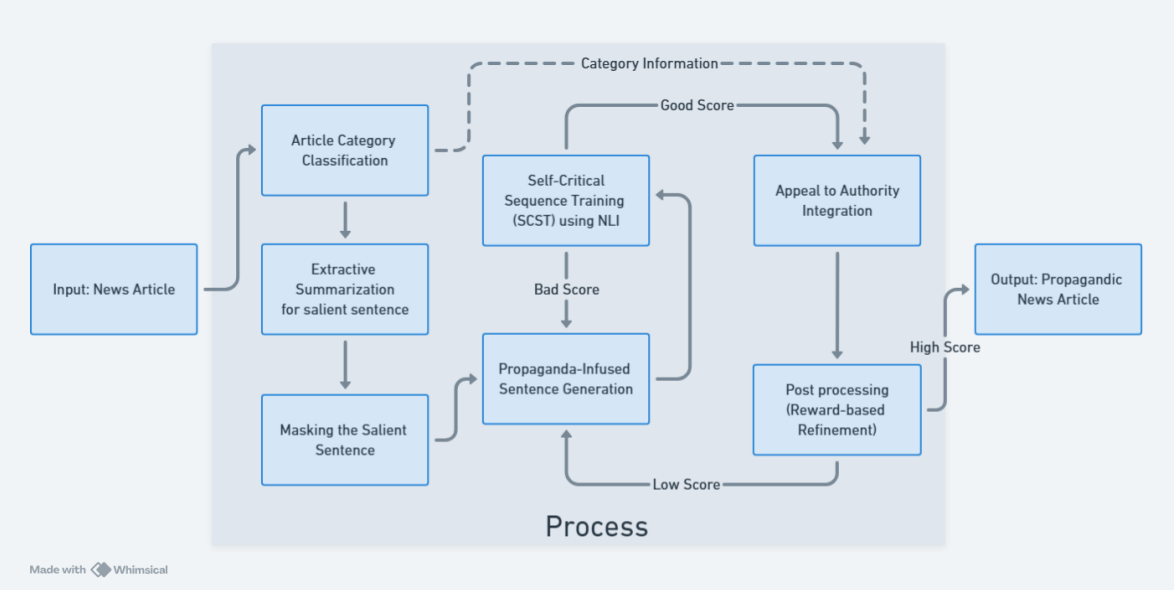

FIGURE 3.3: Disinformation generation pipeline: from input article to final output.

CHAPTER 4

Methodology – Part II: Propaganda Detection Model

This chapter presents the design and implementation of a hybrid BERT+GAT model for binary classification of propaganda in news articles. The methodology covers key stages including data preprocessing, feature extraction, graph construction, model architecture, and the training strategy. Each component is designed to capture both the deep contextual semantics of individual articles and the relational patterns across them, forming a comprehensive detection framework.

4.1 Data Preprocessing

The preprocessing stage begins with the integration of two data sources: real news articles collected from reliable outlets and synthetic disinformation articles generated using the PropDataSet pipeline. These two sources are combined to form a balanced dataset that represents both genuine and propagandistic content. Each article is assigned a binary label: `1` for propaganda (synthetically generated using rhetorical strategies) and `0` for non-propaganda (real, unaltered news). This binary format forms the basis of the classification task and ensures the model can distinguish between manipulated and authentic narratives.

Once the dataset is constructed, each sample undergoes a series of preprocessing steps to prepare it for both contextual and structural analysis:

- **Text Cleaning and Normalization:** All textual data is cleaned by removing irrelevant characters, HTML tags, URLs, and special symbols. The text is normalized by converting it to lowercase to ensure uniform token representation.
- **Sentence Segmentation and Tokenization:** Articles are segmented into sentences using NLP libraries such as spaCy or NLTK. Tokenization is performed to convert the sentences into a sequence of tokens compatible with BERT input formatting.
- **Label Encoding:** The combined dataset is labeled such that synthetic propaganda articles are marked as `1`, and real news articles are marked as `0`. These binary labels are later used during training and evaluation.
- **Length Handling for BERT Input:** Given BERT's input limitation of 512 tokens, longer articles are chunked into smaller segments. Each segment is padded or truncated as needed, and special tokens ([CLS], [SEP]) are inserted to mark the sequence boundaries.
- **Preparation for Feature Extraction:** The cleaned and tokenized samples are then fed into a feature extraction module to compute both contextual embeddings and linguistic attributes for downstream graph-based modeling.
- **Data Storage and Indexing:** To enable reproducibility and efficient batch loading, all preprocessed data including tokenized text, labels, and metadata—are stored in structured format like Pickle.

This preprocessing pipeline ensures that both real and synthetic articles are uniformly processed and encoded, setting a consistent foundation for feature extraction and graph construction in the subsequent steps of the detection framework.

4.2 Feature Extraction

To represent each news article effectively for the downstream classification task, two complementary types of features are extracted: contextual embeddings derived from BERT and domain-specific linguistic features. This hybrid representation captures both the nuanced semantics of the text and rhetorical patterns commonly associated with propaganda.

4.2.1 BERT Embeddings

Contextual embeddings are extracted using the `bert-base-uncased` model [11], a widely adopted Transformer-based language model pre-trained on large-scale English corpora. BERT excels at encoding deep semantic relationships by leveraging bidirectional attention mechanisms, making it well-suited for detecting subtle linguistic cues in propaganda-laden content.[26]

To address the input length constraint of BERT (maximum of 512 tokens), a hierarchical embedding strategy is employed:

- Each article is first tokenized using the BERT tokenizer and divided into contiguous chunks of up to 510 tokens, reserving space for special tokens `[CLS]` and `[SEP]`.
- Each chunk is independently passed through BERT, and the `[CLS]` token embedding—representing the entire sequence—is extracted from the final hidden layer. This results in a 768-dimensional embedding per chunk.
- For articles spanning multiple chunks, the individual `[CLS]` embeddings are averaged to produce a single, fixed-size 768-dimensional representation for the entire article.

To manage GPU memory usage and ensure computational efficiency, embeddings are processed in mini-batches (e.g., 16 samples per batch) and saved to disk in serialized formats for reuse during training and evaluation phases.

4.2.2 Linguistic Features

In addition to contextual embeddings, five task-specific linguistic features are extracted to capture emotional tone, stylistic traits, and rhetorical markers frequently associated with propaganda techniques [30]. These handcrafted features supplement the neural embeddings and provide the model with additional domain-relevant signals:

- **Sentiment Polarity:** The overall sentiment of the article is computed using TextBlob [31], which assesses lexical tone based on predefined polarity scores. This helps identify emotionally charged language, a hallmark of propaganda.
- **Readability Score:** The Flesch Reading Ease metric, computed via the `textstat` library, is used to assess textual complexity [32]. Simpler sentence structures may indicate an intent to persuade or manipulate less critically engaged audiences.
- **Positive Word Count:** The number of positively connoted words (e.g., *good*, *amazing*, *wonderful*) is tallied to identify articles that use optimistic framing to appeal emotionally to readers.
- **Negative Word Count:** Similarly, words with negative sentiment (e.g., *bad*, *horrible*, *terrible*) are counted, providing insight into fear-based or adversarial rhetoric common in propagandistic texts [33].
- **Propaganda Keyword Frequency:** A curated list of domain-specific terms (e.g., *fake*, *hoax*, *bias*, *mislead*, *conspiracy*) is used to count explicit signals of misleading or manipulative content [8].

All linguistic features are standardized using z-score normalization (zero mean, unit variance) to ensure compatibility with BERT's continuous feature space. The final representation for each article is obtained by concatenating the 768-dimensional BERT embedding with the 5 linguistic features, resulting in a 773-dimensional composite feature vector.

This combined feature set serves as input to the graph construction and classification modules, providing both semantic richness and task-relevant cues for effective propaganda detection.

4.3 Graph Construction

To capture inter-document relationships and enhance the model's ability to detect coordinated patterns of disinformation, an undirected graph is constructed in which each node represents a news article and each edge encodes the similarity between articles. This graph structure enables the application of Graph Attention Networks (GAT), which rely on neighborhood connectivity to propagate and attend to contextual signals across nodes.

4.3.1 Graph Structure and Motivation

The graph is designed as an undirected and weighted structure, where the symmetry of edges reflects the bidirectional nature of content similarity, i.e., if Article A is similar to Article B, then B is equally similar to A. This undirected representation is suitable for modeling mutual relationships such as shared vocabulary, semantic overlap, or rhetorical framing, which are crucial for identifying clusters of propaganda-laden content [34].

4.3.2 Similarity Computation

Edges in the graph are established based on a combined similarity score between article pairs. This similarity is computed using two complementary approaches:

- **TF-IDF Similarity:** Each article is transformed into a TF-IDF vector using a domain-specific vocabulary focused on propaganda-related terminology (e.g., *fake*, *hoax*, *conspiracy*, *bias*). Cosine similarity is then computed between all article pairs. For each article, the top $k = 10$ most similar neighbors are retained using a k-nearest neighbors (k-NN) approach. TF-IDF emphasizes explicit lexical overlap, which is often indicative of shared rhetorical strategies in propaganda detection [35].
- **BERT Similarity:** Cosine similarity is also computed between the 768-dimensional BERT embeddings of article pairs. As BERT captures deep contextual semantics,

this method is effective at identifying related content even when the surface-level vocabulary differs. Again, a k-NN strategy with $k = 10$ is applied to retain the most relevant semantic neighbors [11].

- **Combined Similarity:** To leverage both lexical and contextual information, the final similarity score is computed as a weighted combination of the two methods:

$$\text{combined\_sim}(A, B) = 0.7 \times \text{TF-IDF\_sim}(A, B) + 0.3 \times \text{BERT\_sim}(A, B)$$

Empirical testing across multiple datasets demonstrated that a 0.7 weighting on TF-IDF yielded improved model performance—particularly in attending to domain-relevant signals—while the 0.3 contribution from BERT enhanced semantic generalization. This blend proved effective in improving the F1-score and overall classification accuracy for propaganda detection [36].

4.3.3 Edge Creation and Thresholding

An edge is added between two articles if their combined similarity exceeds a threshold of 0.5. This threshold ensures that only significantly related samples are connected, preserving strong semantic or rhetorical ties while avoiding noise from weak or incidental similarities. The weight of the edge corresponds to the combined similarity value, allowing the GAT model to assign higher attention scores to more relevant neighbors during training [12].

4.3.4 Graph Representation and Storage

The final graph is encoded using the PyTorch Geometric library, a widely adopted framework for graph-based deep learning. The graph is stored as a `Data` object with the following components:

- `x`: Node features, including the 773-dimensional vector for each article (768 from BERT + 5 linguistic features).

- `edge_index`: A tensor representing the connectivity of the graph, i.e., pairs of node indices that are connected.
- `edge_attr`: Edge weights corresponding to the combined similarity values.
- `y`: Binary class labels for each node (0 for non-propaganda, 1 for propaganda).

This graph structure enables the GAT model to operate effectively, using attention-based neighborhood aggregation to propagate relevant information across the document space. It plays a central role in facilitating relational learning and detecting patterns indicative of coordinated disinformation campaigns.

4.4 Model Architecture

To effectively combine deep semantic understanding with relational learning, a hybrid model architecture is proposed that integrates BERT-based contextual embeddings with a Graph Attention Network (GAT). The design allows the model to leverage both the local textual features of individual news articles and the global structural dependencies across the document graph. In addition to the hybrid BERT+GAT configuration, standalone BERT and GAT models are implemented and used in ablation studies to assess the individual contribution of each component [37]. The hybrid model consists of three core components:

4.4.1 GAT Component

The GAT module learns graph-level representations by performing attention-based aggregation over neighboring nodes. It operates on the 773-dimensional feature vectors (768 from BERT and 5 linguistic features) extracted for each node. The GAT module includes:

- **First GAT Layer:** This layer employs 4 parallel attention heads, each projecting the input node features to 64 dimensions. The outputs of all heads are

concatenated to form a 256-dimensional vector per node. This multi-head attention mechanism allows the model to learn different perspectives of neighbor relevance [38].

- **Second GAT Layer:** A single attention head further processes the 256-dimensional vector into a unified 64-dimensional graph-aware representation. This layer refines the learned embeddings by aggregating higher-order neighborhood information.
- **Regularization:** Dropout with a rate of 0.1 is applied after each GAT layer to mitigate overfitting, and Exponential Linear Unit (ELU) activations are used to introduce non-linearity and improve learning dynamics [39].

4.4.2 Fusion Mechanism

To integrate local semantic and global relational features, the original node features (773-dimensional) are concatenated with the 64-dimensional output from the GAT component. The resulting 837-dimensional vector is passed through a fully connected (linear) layer of equal size, followed by ELU activation and a dropout layer (rate = 0.1). This fusion strategy enhances the model's capacity to jointly reason about an article's internal content and its contextual placement in the document network [36].

4.4.3 Output Layer

The fused representation is passed through a final linear projection layer that maps the 837-dimensional vector to a 2-dimensional output. A softmax activation function is applied to obtain class probabilities for binary classification (propaganda vs. non-propaganda). This setup aligns with standard practice in binary classification tasks where probabilistic confidence is beneficial for downstream interpretation and threshold-based decision-making [40].

To facilitate model interpretability, the attention weights from the first GAT layer are retained during inference. These weights reflect the relative importance assigned to each neighboring node and can be visualized to examine how the model identifies

influential documents in the graph. This insight is particularly valuable in the context of propaganda detection, where relational influence and network-level behavior often play a critical role [12].

4.5 GAT Pre-training

To enhance the relational understanding of the model before integrating textual semantics, the Graph Attention Network (GAT) component is pre-trained independently on the constructed graph. This pre-training phase spans 100 epochs and is designed to help the GAT layers capture intrinsic structural patterns and neighborhood dependencies, which are essential for relational reasoning in propaganda detection tasks [12].

During pre-training, the node feature inputs consist of the original 773-dimensional vectors (BERT embeddings concatenated with linguistic features), and supervision is provided using binary class labels. The AdamW optimizer is employed with a learning rate of 5×10^{-4}, weight decay of 1×10^{-4}, and a cross-entropy loss function [41]. Once trained, the GAT layer weights are saved and used to initialize the corresponding components in the full BERT+GAT architecture. This transfer of learned graph representations helps the final model converge faster and generalize better [42].

4.6 Training the BERT+GAT Model

The complete hybrid model is trained end-to-end, combining both the pre-trained GAT and the BERT-based contextual embeddings with task-specific fusion and output layers. The training process is carefully tuned to optimize performance while minimizing overfitting and training instability.

- **Optimizer:** The AdamW optimizer is used with differential learning rates: 1×10^{-4} for pre-trained GAT parameters (to fine-tune gradually) and 1×10^{-3} for newly initialized layers (fusion and output), to allow more aggressive updates where necessary [41]. Weight decay is set to 1×10^{-4} to regularize the model and reduce overfitting risks [43].

- **Learning Rate Scheduler:** A cosine annealing scheduler is employed over a maximum of 300 epochs, enabling smooth decay of the learning rate and facilitating stable convergence during training [44].
- **Training Procedure:** Training proceeds for up to 300 epochs. Early stopping is applied if validation loss does not improve for 30 consecutive epochs (patience = 30), thereby preventing overfitting [45]. Gradient accumulation is performed over 4 steps to handle large batches, and gradient clipping (maximum norm = 1.0) is used to stabilize training, particularly important in graph-based deep learning scenarios [46].
- **Loss Function:** Binary cross-entropy loss is used as the objective function, consistent with binary classification settings and prior work in propaganda detection [40].
- **Memory Management:** CUDA memory is explicitly released after each epoch to avoid memory leaks and ensure scalability on large datasets, in accordance with best practices in GPU-based model training [47].

Figure 4.1 provides a high-level overview of the proposed detection framework. It illustrates the integration of BERT-based embeddings for linguistic understanding with a GAT-based graph structure for relational learning. The architecture enables the system to simultaneously exploit both local textual features and global graph-level patterns, enhancing its ability to identify both standalone and coordinated propaganda content.

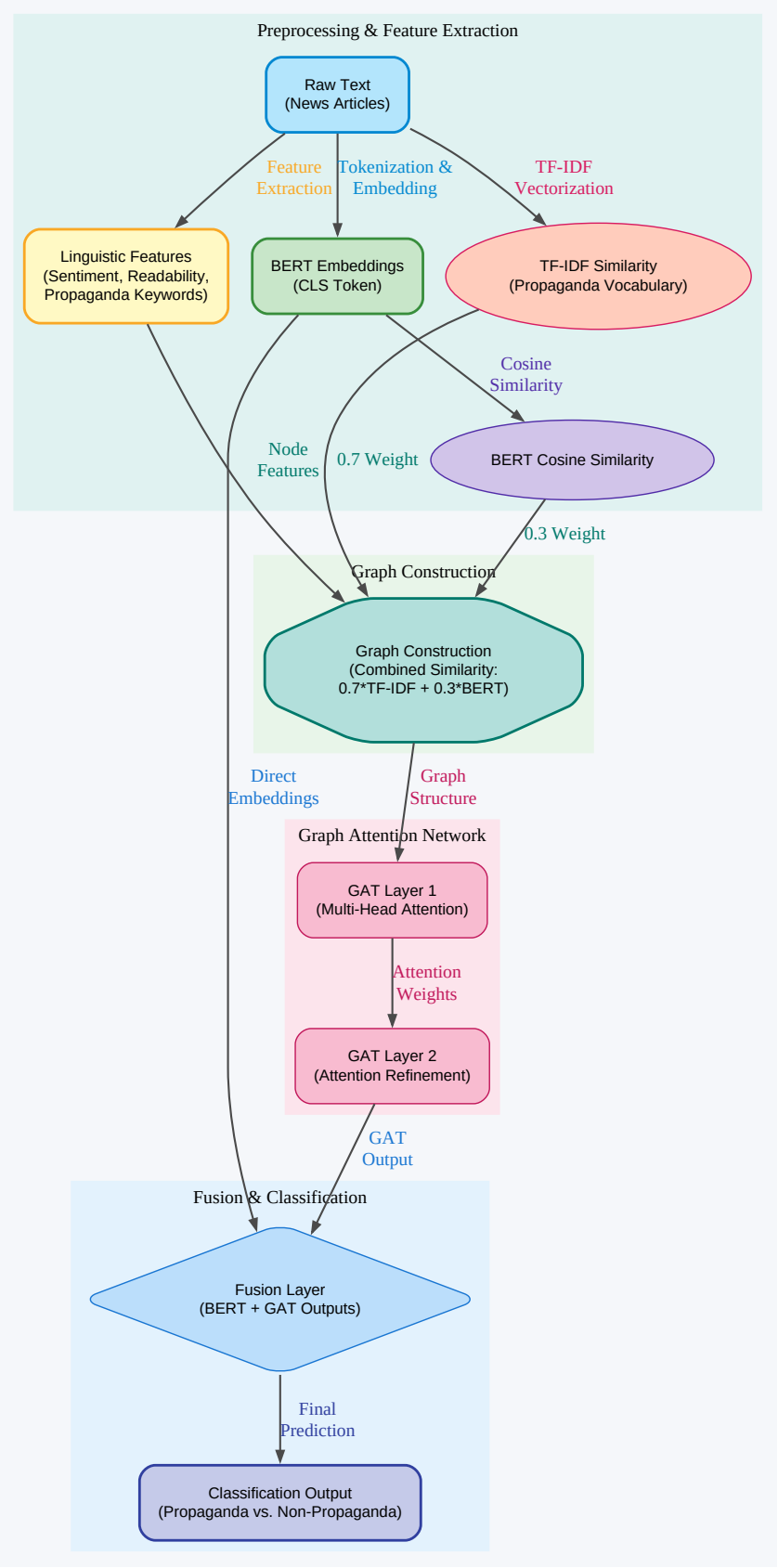

FIGURE 4.1: Flowchart of Detection Approach

CHAPTER 5

Experimental Results and Discussions

5.1 Evaluation of Generated Disinformation Content

This section evaluates the performance of our disinformation generation model and and its generated content. We focus on key metrics such as coherence, factual accuracy, and toxicity to assess the quality, credibility, and psychological impact of the generated content. Comparisons with other state-of-the-art datasets, PropaNews[25] and Grover[19], further highlight our datasets's strengths and limitations.

5.1.1 Evaluation Metrics

The effectiveness of disinformation relies on its ability to appear plausible while distorting facts and influencing opinions. We evaluate the model's output using three critical metrics:

5.1.1.1 Coherence and Plausibility (Entailment Score)

The coherence of the generated disinformation is assessed using the **Entailment Score**, which measures the logical relationship between the original article (premise) and the generated disinformation (hypothesis). This score is calculated using a Natural Language Inference (NLI)[1] model, which classifies the relationship between the premise and hypothesis as:

- **Entails:** The hypothesis logically follows the premise, maintaining coherence and plausibility.
- **Contradicts:** The hypothesis directly contradicts the premise, introducing distortions or biases that are characteristic of disinformation.
- **Neutral:** The hypothesis does not clearly relate to the premise, offering a neutral stance.

Interpretation:

- **High Entailment Score:** Indicates strong alignment with the premise, which may result in content that is too factual to qualify as effective disinformation.
- **Moderate Entailment Score:** Reflects subtle contradictions or distortions, creating misleading yet plausible content.
- **Low Entailment Score:** Suggests significant deviations from the premise, producing disinformation that is overtly fabricated or less believable.

Example:

- **Premise:** "The new climate change legislation has been widely praised by environmentalists."
- **Moderate Entailment Hypothesis:** "Environmentalists have shown mixed reactions, with some praising the legislation and others criticizing it for being ineffective."

[1] https://huggingface.co/FacebookAI/roberta-large-mnli

- **Low Entailment Hypothesis:** "Environmentalists have condemned the new climate change legislation as harmful."

The first example introduces mild contradictions, while the second creates a stark false narrative. A moderate entailment score is ideal for generating convincing propaganda.

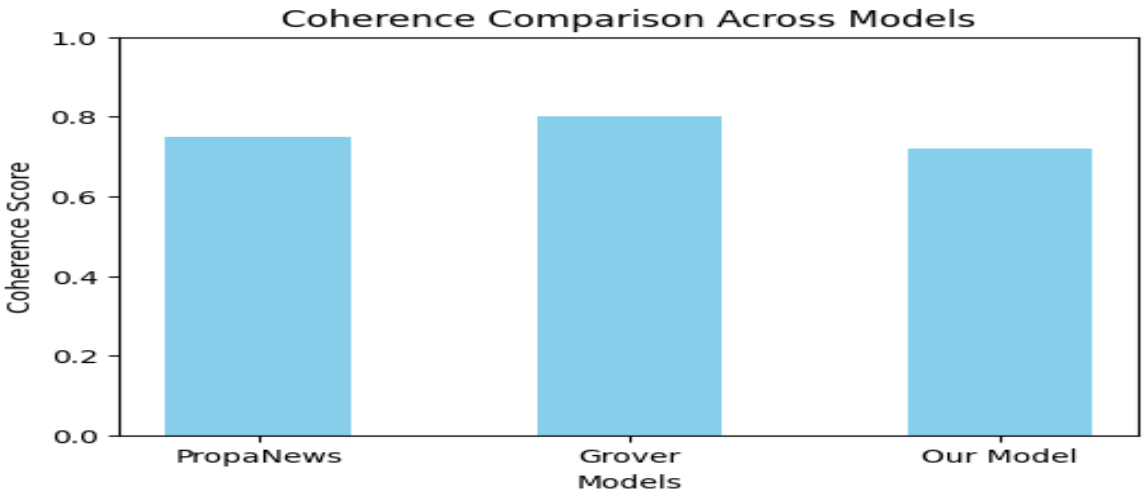

FIGURE 5.1: Coherence Score Comparison

5.1.1.2 Factual Accuracy (ClaimBuster Assessment)

Evaluating the factual accuracy of generated content is crucial in understanding its ability to mislead while maintaining a sense of credibility. For this purpose, we utilized **ClaimBuster**[2], an advanced automated fact-checking tool designed to identify and assess claims made in public discourse.ClaimBuster combines machine learning and natural language processing to analyze large volumes of text and pinpoint statements that can be fact-checked. It is widely used in political and journalistic contexts for verifying the accuracy of claims made in news articles, speeches, and social media posts.

How ClaimBuster Works:

- **Claim Identification:** ClaimBuster scans content such as articles, political debates, or social media discussions to detect specific statements that appear

[2]https://idir.uta.edu/claimbust

factual and verifiable. For instance, a sentence like, "The unemployment rate has dropped to its lowest in a decade," would be flagged for verification.

- **Contextual Analysis:** The tool analyzes the context of the detected claim to determine whether it is specific enough to be verified or too ambiguous for evaluation. For example, broad statements like "Taxes are too high" would be categorized as non-verifiable due to their subjective nature.
- **Verification Process:** Once a claim is identified, ClaimBuster cross-references it against a curated database of reliable sources, including government records, research studies, and verified fact-checking repositories. This ensures that the evaluation is grounded in credible evidence.
- **Results and Scoring:** Each claim is assessed for its verifiability and given a confidence score. This score reflects the reliability of the supporting evidence and the likelihood that the claim is accurate. Claims with higher scores are considered more factual, while lower scores suggest potential misinformation.

For our disinformation model, we used ClaimBuster to analyze generated content for its alignment with factual accuracy. By submitting statements produced by the model, we were able to determine whether the disinformation embedded subtle distortions of truth or blatant fabrications.

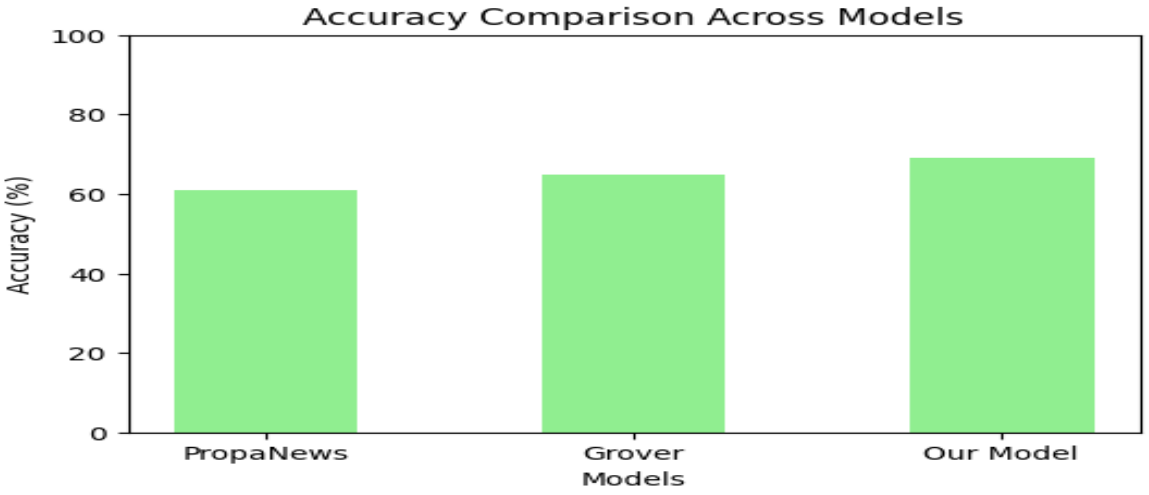

FIGURE 5.2: Factual Accuracy Comparison

5.1.1.3 Toxicity Assessment (Perspective API Analysis)

Toxicity plays a significant role in shaping how readers perceive and engage with disinformation. Content that includes harmful or offensive language often provokes strong emotional reactions, fostering division or amplifying ideological biases. To assess the toxicity of our model's generated content, we utilized the **Perspective API**[3], a tool developed specifically to evaluate the potential impact of language in conversations.

The API assigns a score between 0 and 1 for each attribute, indicating the likelihood that the analyzed text contains that characteristic. For example:

- A score close to **1** for the Toxicity attribute suggests that the content is highly inflammatory or divisive.
- A lower score (closer to **0**) indicates more neutral or non-provocative language.

These scores help quantify the emotional impact and potential harm of the content, making it easier to identify and address inflammatory text. We employed the Perspective API to analyze the toxicity of content generated by our disinformation model. By evaluating the text for attributes like general toxicity, insults, and identity attacks, we were able to gain insights into how inflammatory or harmful the generated content might be.

5.1.2 Comparison of Datasets

To evaluate the effectiveness of our disinformation generation model, we compared its generated content against two prominent datasets: PropaNews[25] and Grover[19]. Below are the results:

PropaNews exhibited a moderate coherence score ranging between 0.70 and 0.75, indicating that its generated disinformation maintains a reasonable logical flow with occasional biases. However, its factual accuracy was relatively low at 61%, with notable

[3]https://perspectiveapi.com/

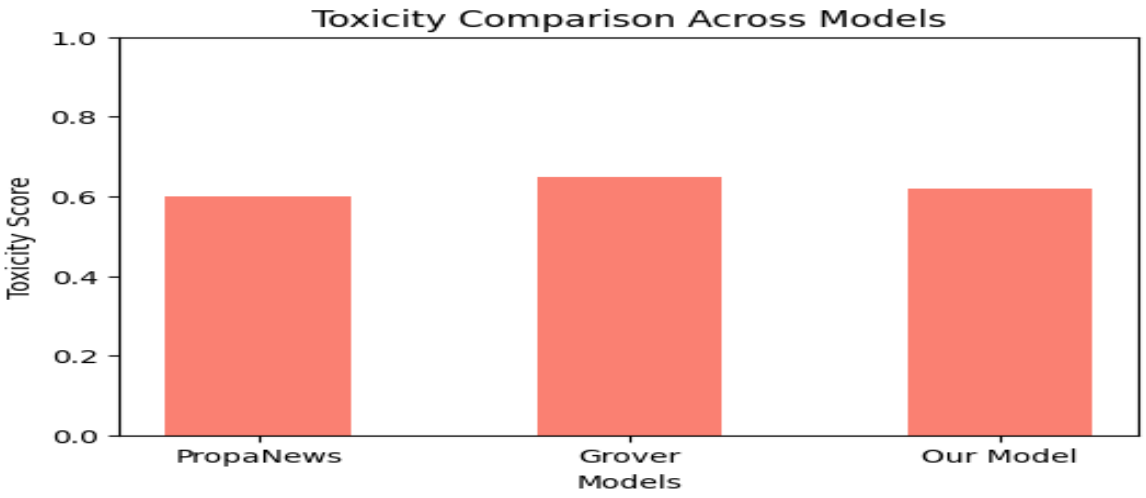

FIGURE 5.3: Toxicity Comparison

Model	Coherence	Factual Accuracy	Toxicity Score
PropaNews	0.75 (Moderate)	61%	0.60
Grover	0.80 (High)	65%	0.65
Our Model	0.72 (Moderate)	69%	0.62

TABLE 5.1: Performance Comparison Across Metrics

instances of distortion in its generated content. Its toxicity score stood at 0.60, reflecting moderately inflammatory language designed to provoke some degree of emotional response.

Grover achieved the highest coherence score at 0.80, producing highly plausible and logically consistent content. Despite this, its factual accuracy was only marginally better than PropaNews at 65%, suggesting that it prioritizes logical flow over truthfulness. Its toxicity score of 0.65 was the highest among the three models, indicating that Grover often employed provocative language to engage readers.

Our model demonstrated a coherence score of 0.72, which aligns closely with PropaNews but remains slightly below Grover. This result reflects the model's intentional inclusion of subtle contradictions and biases, enhancing its ability to mislead while still appearing logical. In terms of factual accuracy, our model outperformed the others, achieving a score of 69%. This highlights its ability to generate disinformation that feels believable by carefully balancing factual elements with misleading narratives. Its toxicity score of

0.62 strikes a middle ground, reflecting the use of moderately inflammatory language without overreliance on harmful rhetoric.

The data highlights the distinct strengths and trade-offs of each model. While Grover excels in producing coherent and logical text, its relatively low factual accuracy and high toxicity score suggest a tendency to generate highly provocative disinformation that sacrifices truthfulness. PropaNews, on the other hand, delivers less coherent content with lower factual accuracy, though it employs slightly less inflammatory language. Our model stands out by striking a balance: it maintains moderate coherence, the highest factual accuracy, and controlled toxicity levels. This balance ensures the content is misleading enough to serve as effective propaganda while minimizing the risk of being flagged for excessively inflammatory language, making it a more strategic tool for generating persuasive disinformation.

5.2 Evaluation of Propaganda Detection Model

This section presents a comprehensive evaluation of the proposed BERT+GAT propaganda detection model across three diverse datasets: PropDataSet, WELFake, and MisInfoSet. These datasets encompass a broad range of disinformation types, sources, and linguistic styles, providing a robust foundation for assessing model effectiveness. The experiments focus on evaluating classification accuracy, robustness, and generalization across both synthetic and real-world scenarios, enabling a thorough comparison of the hybrid model against several state-of-the-art baselines.

5.2.1 Datasets for Experimentation

To test the effectiveness and adaptability of the proposed detection framework, experiments are conducted on the following three datasets:

- **PropDataSet** – A synthetically generated dataset created as part of this research, embedding controlled propaganda techniques into real news articles.

- **WELFake** – A large-scale, publicly available benchmark dataset for fake news detection, containing a variety of both human- and AI-generated misinformation.
- **MisInfoSet** – A curated collection of articles from mainstream and fringe media, including verified propaganda instances from international monitoring bodies.

Each dataset provides unique challenges, allowing the proposed model to be evaluated across different domains, textual styles, and disinformation strategies.

5.2.1.1 PropDataSet: A Synthetic Propaganda Dataset

PropDataSet was developed using the disinformation generation pipeline described in Chapter 3. It simulates real-world propaganda by embedding rhetorical manipulation techniques such as appeal to authority, flag-waving, and loaded language into authentic news articles. The dataset is binary-labeled: articles containing propaganda are marked as `1`, while unaltered, fact-based news are labeled as `0`.

Designed to address the scarcity of high-quality labeled data, PropDataSet offers a controlled, scalable, and balanced environment for training models to detect both factual distortion and persuasive framing. Its construction allows for fine-grained learning of propaganda cues in a way that complements traditional fake news datasets.

5.2.1.2 WELFake: A Large-Scale Fake News Dataset

WELFake [48] is one of the most comprehensive datasets in the field of fake news detection, comprising approximately 72,000 news articles from multiple sources. It aggregates content from platforms like Kaggle, McIntire, BuzzFeed, and Reuters, featuring a broad mix of authentic and fabricated news.

WELFake is particularly useful for evaluating general fake news detection models due to its scale and diversity. The inclusion of both human-authored and machine-generated misinformation makes it suitable for assessing the robustness of propaganda detection systems against various forms of disinformation.

5.2.1.3 MisInfoSet: A Diverse News and Propaganda Dataset

MisInfoSet [49] is a curated dataset that integrates content from mainstream news outlets, fringe platforms, and verified disinformation cases. It provides a real-world snapshot of how misinformation and propaganda are disseminated across different channels. The dataset includes:

- Articles from credible sources such as Reuters [50], The New York Times [51], and The Washington Post [52].
- Misinformation from outlets like Redflag Newsdesk [53], Breitbart [54], and Truth Broadcast Network [55].
- Disinformation cases identified by the EUvsDisinfo project [56], focused on pro-Kremlin propaganda narratives.
- Historical fake news samples from Ahmed et al. (2017) [57], which use n-gram-based feature extraction and machine learning for fake news classification.

MisInfoSet's diversity allows the evaluation of how well models generalize across domains, textual styles, and ideologically different sources. It tests the capacity of the detection framework to identify subtle propaganda even in noisy, real-world data.

5.2.2 Dataset Utilization in Various Model Training

To ensure a fair and comprehensive evaluation of our proposed detection framework, each of the three datasets—PropDataSet, WELFake, and MisInfoSet—is used to independently fine-tune and test a suite of baseline and hybrid models. This controlled experimental setup allows for a robust comparison of model effectiveness under varying disinformation characteristics.

The models evaluated include:

1. **BERT:** A fine-tuned instance of the base BERT model trained for binary classification of propaganda at the article level [11].

2. **RoBERTa:** A variant of BERT that incorporates robust pre-training and optimization strategies to enhance language representation capabilities [58].

3. **GCN:** A Graph Convolutional Network trained on graph-structured data to model relational patterns among articles through local neighborhood aggregation [59].

4. **GAT:** A Graph Attention Network that extends GCN by using attention mechanisms to assign varying importance to different neighboring nodes during message passing [12].

5. **BERT + GAT (Proposed Model):** A hybrid architecture combining BERT for deep semantic feature extraction with GAT for structural and relational learning, aiming to bridge the gap between content-level and network-level disinformation patterns.

Each model is trained and evaluated separately on all three datasets, using consistent preprocessing, training parameters, and evaluation metrics to ensure comparability. The following section presents the results and analysis.

5.2.3 Model Performance and Analysis

5.2.3.1 Performance Overview

Table 5.2 presents a quantitative comparison of model performance across the three datasets, using standard classification metrics—Accuracy, Precision, Recall, and F1-score. As shown, the proposed BERT+GAT model achieves the highest performance across all metrics and datasets, demonstrating the advantage of combining contextual and structural representations for propaganda detection.

- On **PropDataSet**, the BERT+GAT model attains a remarkable accuracy of **99.87%**, indicating its effectiveness in detecting synthetically embedded propaganda techniques.

TABLE 5.2: Performance Comparison of Techniques Across Misinformation Detection Datasets for various models

Technique	PropDataSet				Misinfo				WELFake			
	Acc	**F1**	**Rec**	**Prec**	**Acc**	**F1**	**Rec**	**Prec**	**Acc**	**F1**	**Rec**	**Prec**
RoBERTa	0.9356	0.9384	0.9274	0.9365	0.9294	0.9275	0.9034	0.9530	0.9129	0.9141	0.9269	0.9016
GCN	0.9744	0.9703	0.9688	0.9718	0.8166	0.8180	0.8246	0.8116	0.8197	0.8168	0.8040	0.8301
BERT	0.9247	0.9248	0.9250	0.9247	0.9020	0.9018	0.9000	0.9036	0.8997	0.9000	0.9023	0.8977
GAT	0.9785	0.9787	0.9756	0.9818	0.8789	0.8764	0.8589	0.8946	0.8577	0.8572	0.8543	0.8602
BERT-GAT Fusion	0.9987	0.9988	0.9994	0.9981	0.9591	0.9595	0.9669	0.9522	0.9586	0.9587	0.9611	0.9562

- On **MisInfoSet**, it achieves **95.91%** accuracy, outperforming all other baselines in handling heterogeneous, real-world disinformation.
- On **WELFake**, the hybrid model maintains a high accuracy of **95.86%**, showcasing its robustness even when tested against large-scale, noisy datasets.

The results are visualized in Figure 5.4, further confirming that the hybrid model consistently outperforms standalone BERT, RoBERTa, and graph-based models across all datasets.

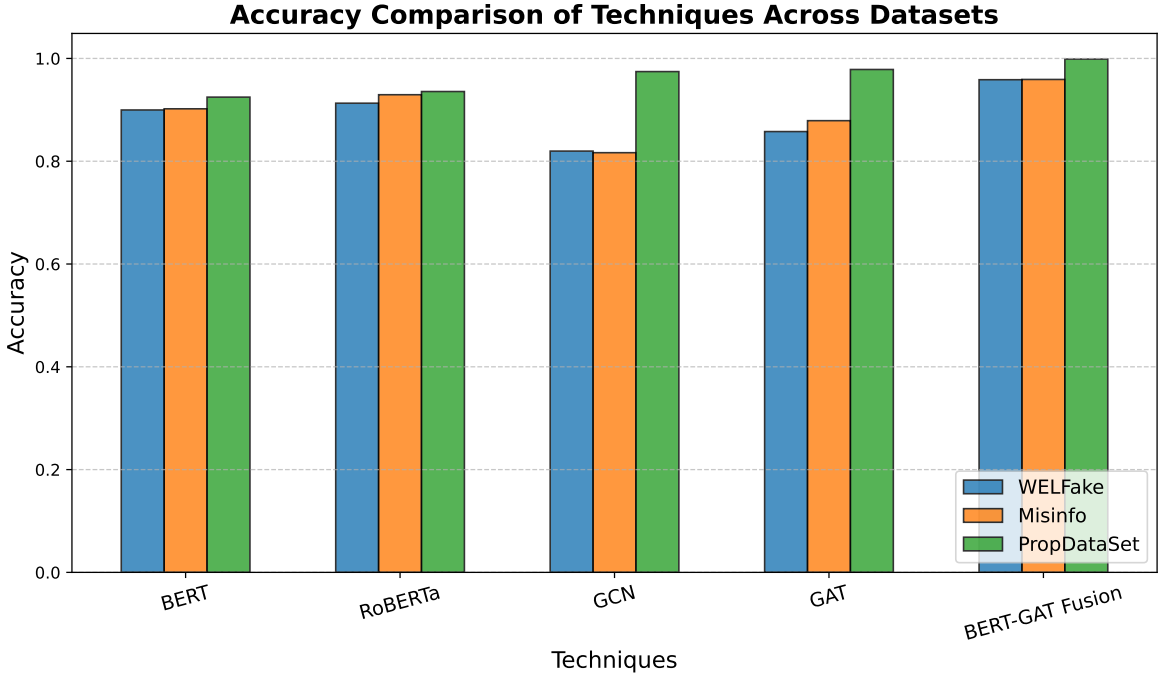

FIGURE 5.4: Accuracy Comparison of Techniques Across Datasets for various models

5.2.3.2 Cross-Dataset Generalization

The generalization ability of the BERT+GAT model is also assessed through cross-dataset evaluations, where a model trained on one dataset is tested on another. Table 5.3 and Figure 5.5 illustrate that the hybrid model trained on PropDataSet generalizes well to both WELFake and MisInfoSet, indicating that the synthetic dataset effectively captures transferable features of propagandistic content.

TABLE 5.3: Cross-Domain performance of BERT-GAT Model Trained on PropDataSet

Dataset	Accuracy	Precision	Recall	F1
Misinfo	0.9670	0.9600	0.9700	0.9650
WELFake	0.9650	0.9620	0.9680	0.9650
PropDataSet	0.9987	0.9988	0.9994	0.9981

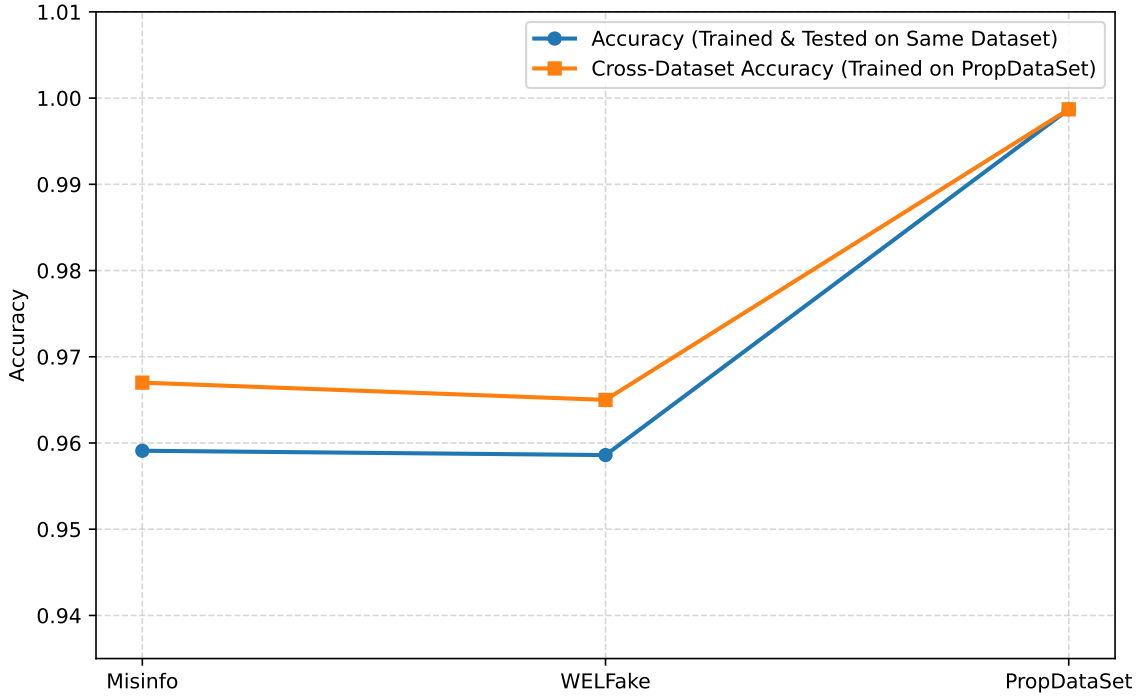

FIGURE 5.5: Cross-Domain Accuracy of BERT-GAT model trained on PropDataSet

5.2.3.3 Dataset-Specific Insights

PropDataSet: As a synthetically generated dataset incorporating well-defined propaganda techniques, PropDataSet proves highly effective for training models to detect subtle and structured disinformation. The near-perfect performance by the BERT+GAT model demonstrates the dataset's quality and its utility for AI-driven propaganda detection research.

MisInfoSet and WELFake: These datasets represent real-world challenges, featuring unstructured, diverse misinformation. Traditional models such as BERT and RoBERTa show moderate success, but the hybrid BERT+GAT consistently achieves superior performance. This suggests that relational modeling, when combined with deep semantic features, significantly enhances a model's ability to detect varied and noisy disinformation patterns in practical settings.

5.3 Key Insights from the results

- **Our PropDataSet dataset greatly improves model learning to facilitate improved detection of AI-generated propaganda.**
- **Hybrid models (BERT-GAT Fusion) consistently report the highest accuracy, as shown in Tables 5.2 and 5.3, demonstrating the efficacy of integrating transformers with graph-based methods.**
- **Graph-based models excel at detecting structured propaganda but struggle with highly diverse or unstructured misinformation.**
- **Transformer models are powerful, but their performance improves significantly when enhanced with graph-based relational modeling.**
- **The structure of the dataset is very important for model performance, and PropDataSet is a useful resource for the study of misinformation detection.**

CHAPTER 6

Conclusion and Future Work

6.1 Conclusion

This thesis presents a comprehensive framework for improving propaganda detection through two major contributions: the generation of high-quality synthetic disinformation datasets and the development of a hybrid detection model that integrates contextual and relational learning.

In the first phase of the project, we designed a robust pipeline for disinformation generation inspired by real-world propaganda strategies. By incorporating rhetorical techniques such as loaded language, appeal to authority, flag-waving, and fear-based messaging, the system created realistic and diverse propaganda content. The resulting dataset—PropDataSet—was generated by modifying real news articles using controlled semantic alterations and embedding human-like persuasion strategies. This dataset addressed key limitations in existing resources, including annotation bias, limited scale, and lack of rhetorical diversity.

Building on this, the second phase of the research focused on leveraging this synthetic data to train and evaluate a hybrid detection model. The proposed BERT+GAT architecture successfully combined deep semantic understanding from BERT with relational reasoning via Graph Attention Networks. Experimental results on PropDataSet,

WELFake, and MisInfoSet demonstrated that the hybrid model significantly outperformed standalone models like BERT, RoBERTa, GCN, and GAT. Notably, the model generalized well across different datasets, highlighting its robustness in real-world misinformation detection scenarios.

Together, these contributions demonstrate that combining high-quality synthetic training data with hybrid deep learning models offers a promising direction for detecting propaganda more accurately and reliably—especially when real annotated data is scarce or unbalanced.

6.2 Future Work

While the results are promising, there are several opportunities for future improvement and expansion:

- **Expanding Propaganda Techniques:** Increasing the diversity of rhetorical strategies used in synthetic generation, such as name-calling, glittering generalities, or transfer, could enhance the realism and utility of the dataset.
- **Reinforcement and Adversarial Learning:** Integrating techniques such as reinforcement learning or adversarial training could help generate even more contextually accurate and deceptive disinformation samples.
- **Multimodal Propaganda Detection:** Extending the system to analyze not just text but also images, videos, and memes using multi-modal transformers would provide a more complete detection framework suited for modern disinformation.
- **Multilingual and Cross-Domain Support:** Propaganda is not language-specific. Adapting the models to perform effectively across different languages, topics, and platforms using domain adaptation techniques is essential for broader applicability.
- **Temporal and Interaction-Aware Graphs:** Enhancing the graph construction by incorporating temporal information (e.g., article timestamps) and user

interaction data (e.g., shares, retweets) could improve the detection of coordinated disinformation campaigns.

- **Knowledge Graph Integration:** Using knowledge graphs to link entities, claims, and sources could help uncover hidden relationships and detect factual inconsistencies at scale.
- **Deployment at Scale:** For real-world use, the framework must be scalable, fast, and privacy-respecting. Future work could explore cloud-based inference, federated learning, and edge-optimized models to support secure and large-scale deployment.

6.3 Ethical Considerations and Broader Impact

Throughout this project, ethical responsibility has been a core consideration. While the synthetic data generation and detection framework were developed solely for research purposes, we acknowledge the dual-use risks associated with such technologies. Key ethical concerns include:

- **Generation of Toxic or Harmful Content:** Even unintentionally, synthetic models may produce inflammatory or misleading narratives.
- **Misuse of the Generation Pipeline:** The same tools that can be used to train detection systems could potentially be exploited to create propaganda at scale.
- **Bias in Datasets or Models:** Relying on specific sources or linguistic features might introduce unintended bias that affects detection fairness.

To mitigate these risks, we restrict PropDataSet to academic and research purposes only and strongly advocate for transparency and ethical oversight in any deployment of generative models. Future iterations of this research should include regular audits, open access benchmarks, and collaborative efforts with policymakers to establish safeguards.

Ultimately, the goal of this research is to contribute constructively to the fight against misinformation and propaganda, and to promote more reliable and trustworthy information ecosystems.

References

[1] I. Institute For Propaganda Analysis, *How to detect propaganda*. Institute for Propaganda Analysis, 1938.

[2] G. Jowett and V. O'Donnell, *Propaganda and Persuasion*. Sage, 2006.

[3] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," 2017.

[4] Z. Labs and H. Poll, "Study: 86% of people don't fact-check news spotted on social media," *PRWeek*, 2017.

[5] O. U. Press, "54% of indian users rely on social media for factual information," *Gadgets 360*, 2022.

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

[7] A. Barrón-Cedeño, I. Jaradat, G. Martino, and P. Nakov, "Proppy: Organizing the news based on their propagandistic content," *Information Processing Management*, vol. 56, 05 2019.

[8] 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 2017 Conference on Empirical Methods in Natural Language Processing* (M. Palmer, R. Hwa, and S. Riedel, eds.), (Copenhagen, Denmark), pp. 2931–2937, Association for Computational Linguistics, Sept. 2017.

[9] B. D. Horne, W. Dron, S. Khedr, and S. Adali, "Sampling the news producers: A large news and feature data set for the study of the complex media landscape," *CoRR*, vol. abs/1803.10124, 2018.

[10] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, eds.), (Online), pp. 7871–7880, Association for Computational Linguistics, Jul 2020.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), (Minneapolis, Minnesota), Association for Computational Linguistics, Jun 2019.

[12] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," 2018.

[13] H. Rashkin, E. Choi, J. Jang, S. Volkova, and Y. Choi, "Truth of varying shades: Analyzing language in fake news and political fact-checking," pp. 2931–2937, 01 2017.

[14] S. Yoosuf and Y. Yang, "Fine-grained propaganda detection with fine-tuned bert," pp. 87–91, 01 2019.

[15] P. Gupta, K. Saxena, U. Yaseen, T. Runkler, and H. Schütze, "Neural architectures for fine-grained propaganda detection in news," in *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda* (A. Feldman, G. Da San Martino, A. Barrón-Cedeño, C. Brew, C. Leberknight, and P. Nakov, eds.), (Hong Kong, China), pp. 92–97, Association for Computational Linguistics, Nov. 2019.

[16] J. Li, Z. Ye, and L. Xiao, "Detection of propaganda using logistic regression," pp. 119–124, 01 2019.

[17] G. Martino, S. Yu, A. Barrón-Cedeño, R. Petrov, and P. Nakov, "Fine-grained analysis of propaganda in news article," pp. 5640–5650, 01 2019.

[18] G. D. S. Martino, S. Cresci, A. Barron-Cedeno, S. Yu, R. D. Pietro, and P. Nakov, "A survey on computational propaganda detection," 2020.

[19] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, "Defending against neural fake news," 2020.

[20] S. Yu, G. Martino, M. Mohtarami, J. Glass, and P. Nakov, "Interpretable propaganda detection in news articles," pp. 1597–1605, 01 2021.

[21] K. Shu, Y. Li, K. Ding, and H. Liu, "Fact-enhanced synthetic news generation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 13825–13833, 05 2021.

[22] V. Vorakitphan, E. Cabrio, and S. Villata, *PROTECT – A Pipeline for Propaganda Detection and Classification*, pp. 352–358. 01 2022.

[23] P. Vijayaraghavan and S. Vosoughi, "TWEETSPIN: Fine-grained propaganda detection in social media using multi-view representations," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3433–3448, 2022.

[24] K. Sprenkamp, D. G. Jones, and L. Zavolokina, "Large language models for propaganda detection," 2023.

[25] K.-H. Huang, K. McKeown, P. Nakov, Y. Choi, and H. Ji, "Faking fake news for real fake news detection: Propaganda-loaded training data generation," 2023.

[26] S. Chi, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune bert for text classification?," May 2019.

[27] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2009.

[28] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," *Big Data*, vol. 8, pp. 171–188, 06 2020.

[29] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," pp. 7871–7880, 01 2020.

[30] G. D. S. Martino, S. Yu, A. Barrón-Cedeño, R. Petrov, and P. Nakov, "Fine-grained analysis of propaganda in news articles," *arXiv preprint arXiv:1910.02517*, 2019.

[31] S. Loria, P. Keen, M. Honnibal, R. Yankovsky, D. Karesh, and E. Dempsey, "Textblob: simplified text processing," *Secondary TextBlob: simplified text processing*, vol. 3, p. 2014, 2014.

[32] R. Flesch, "A new readability yardstick.," *Journal of applied psychology*, vol. 32, no. 3, p. 221, 1948.

[33] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[34] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[35] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," in *Readings in information retrieval*, pp. 323–328, 1997.

[36] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[37] R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, and J. Leskovec, "Graph convolutional neural networks for web-scale recommender systems," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 974–983, 2018.

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[39] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[40] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*, vol. 1. MIT press Cambridge, 2016.

[41] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[42] W. Hu, B. Liu, J. Gomes, M. Zitnik, P. Liang, V. Pande, and J. Leskovec, "Strategies for pre-training graph neural networks," *arXiv preprint arXiv:1905.12265*, 2019.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[44] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[45] L. Prechelt, "Automatic early stopping using cross validation: quantifying the criteria," *Neural networks*, vol. 11, no. 4, pp. 761–767, 1998.

[46] A. Paszke, "Pytorch: An imperative style, high-performance deep learning library," *arXiv preprint arXiv:1912.01703*, 2019.

[47] D. Podareanu, V. Codreanu, S. Aigner, C. Leeuwen, and V. Weinberg, "Best practice guide-deep learning," *Partnership for Advanced Computing in Europe (PRACE), Tech. Rep*, vol. 2, 2019.

[48] S. Shahane, "Fake news classification dataset," 2022. https://www.kaggle.com/datasets/saurabhshahane/fake-news-classification.

[49] S. Peutz, "Misinformation fake news text dataset (79k)," 2023. https://www.kaggle.com/datasets/stevenpeutz/misinformation-fake-news-text-dataset-79k.

[50] Reuters, "Reuters - breaking news, business, financial and world news," 2025. https://www.reuters.com.

[51] The New York Times, "The new york times - breaking news, us news, world news," 2025. https://www.nytimes.com.

[52] The Washington Post, "The washington post - breaking news, analysis, politics," 2025. https://www.washingtonpost.com.

[53] Redflag News, "Redflag news - alternative news and analysis," 2025. https://redflag.global/news/.

[54] Breitbart News, "Breitbart - conservative news and opinion," 2025. https://www.breitbart.com.

[55] Truth Broadcast Network, "Truth network - conservative talk radio and news," 2025. https://www.truthnetwork.com/.

[56] EUvsDisinfo, "Euvsdisinfo - european union disinformation database," 2025. https://euvsdisinfo.eu.

[57] H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using n-gram analysis and machine learning techniques," in *Proceedings of the International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, pp. 127–138, Oct 2017.

[58] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

[59] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2017.

Appendix I: Graphical Abstract

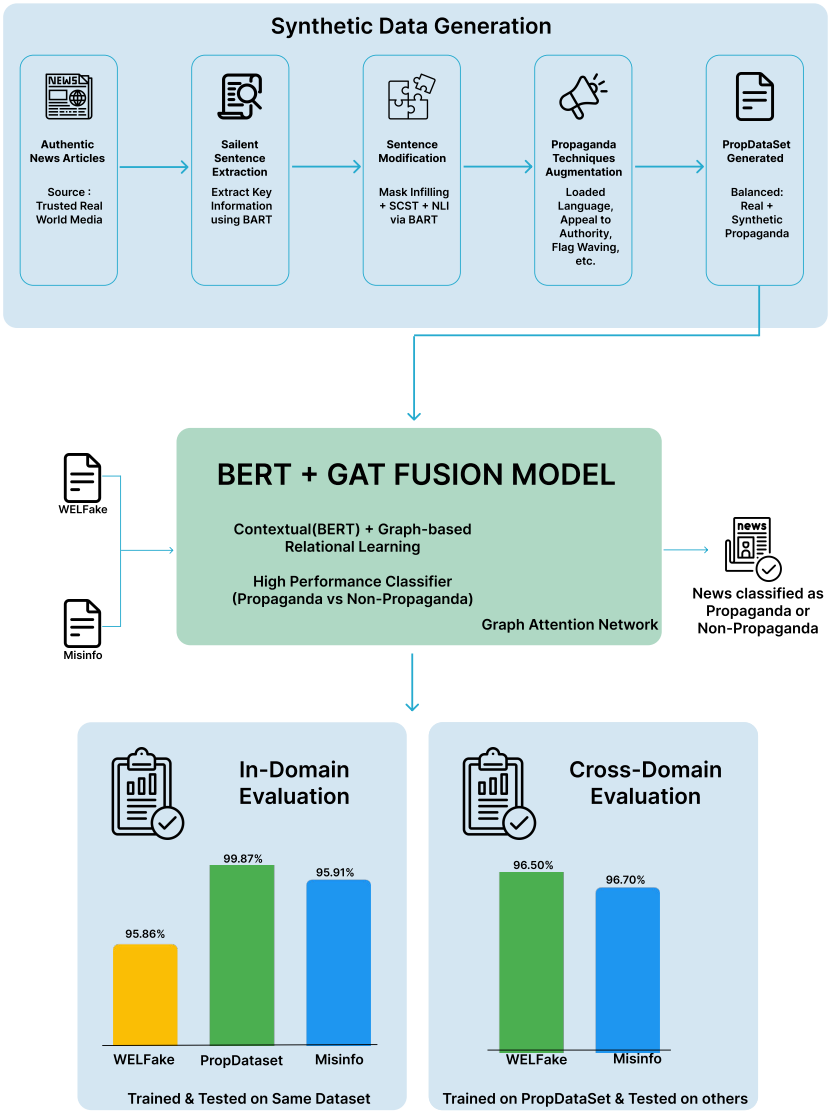

FIGURE 6.1: Graphical Abstract of the Propaganda Detection Framework

Appendix II

Journal under communication

1. **Deep Saikia**, Aryan Kumar Singh, Md Shohan Mia, Dr. Debbrota Paul Chowdhury "**Enhancing Propaganda Detection Using Synthetic Data Informed by Human Disinformation Strategies**", *Expert Systems With Applications,ESWA*, Elsevier.

B.Tech Thesis

ORIGINALITY REPORT

8%	6%	6%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to National Institute of Technology, Silchar Student Paper	2%
2	www.coursehero.com Internet Source	<1%
3	arxiv.org Internet Source	<1%
4	link.springer.com Internet Source	<1%
5	Zhao, Fengxiang. "Building an Extensible, AI-Augmented Ecological Momentary Assessment Platform.", University of Missouri - Columbia, 2024 Publication	<1%
6	dokumen.pub Internet Source	<1%
7	hdl.handle.net Internet Source	<1%
8	assets.researchsquare.com Internet Source	<1%
9	knowledgecommons.lakeheadu.ca Internet Source	<1%
10	Çetinkaya, Yusuf Mücahit. "Bridging AI and Personalization: From Social Media Insights to Targeted Marketing", Middle East Technical University (Turkey)	<1%