AI-DRIVEN FAKE NEWS DETECTION: APPROACHES, TECHNIQUES, EXPERIMENTAL ANALYSIS AND TRENDS

LI ZENG

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

MASTERS OF ARTS

GRADUATE PROGRAM IN INFORMATION SYSTEMS AND TECHNOLOGY
YORK UNIVERSITY
TORONTO, ONTARIO

FEBRUARY 2026

Abstract

The rapid spread of fake news poses a significant challenge to information accuracy. This thesis highlights fake news definitions and characteristics, introducing a taxonomy that categorizes AI-driven detection methods into model-centric and process-centric approaches. We evaluate various approaches ranging from traditional machine learning to trending AI methodologies, focusing on techniques like data augmentation, information extraction, and results explanation.

To re-evaluate classical algorithms, this work provides a detailed analysis of a 2016 U.S. election dataset. By employing fact-checking and advanced data mining, we investigate linguistic characteristics through exploratory data analysis and apply multiple machine learning algorithms for classification.

Experimental results yield valuable insights into the defining characteristics of fake news and demonstrate machine learning's potential to enhance misinformation filtering. Finally, we discuss four main challenges and trends aimed at refining detection accuracy and integrating cutting-edge AI methodologies to combat fake news more effectively.

Acknowledgements

I would like first to express my profound gratitude to my supervisor, Professor Jimmy Huang, whose guidance, support, and expertise have been instrumental in shaping this thesis. His relentless pursuit of excellence and insightful feedback has helped me to explore new avenues of thought and reach greater academic heights. His dedication to my development has truly made this work possible, and I am forever grateful for his trust and encouragement.

I also want to extend my appreciation to the committee members, Professor George Z.H. Zhu and Professor George J. Georgopoulos, for their guidance and suggestions during my Master's defense. Their support and feedback have been invaluable.

Last, but by no means least, I must acknowledge the unwavering support of my parents and loving husband. Their encouragement, patience, and love have sustained me throughout this journey.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background and Significance

News content is disseminated in diverse formats, including text, user comments, videos, audio, etc., but there is a huge challenge to searching for and obtaining the right information: the rampant spread of fake news. The magnitude of this issue is reflected in global public interest and search behaviors. Many scholars point to the 2016 US presidential election as a watershed moment where the viral spread of deceptive content first gained massive notoriety [1, 2]. Since then, keywords such as "fake news"[1] and "Trump fake news"[2] have remained central to public discourse. As illustrated in Figure 1.1 and Figure 1.2, data extracted from Google Trends shows that search interest for these terms typically peaks during major political shifts. The specific phenomenon of "Trump fake news" refers to articles that were intentionally and verifiably false, created to mislead readers during high-stakes election cycles. This trend highlights how fake news has evolved from a technical nuisance into a deliberate strategy used to influence voter sentiment and provoke social panic.

The consequences of fake news are far-reaching. Misinformation has infiltrated critical

[1]https://trends.google.com/trends/explore?date=all&q=fake%20news
[2]https://trends.google.com/trends/explore?date=2015-02-04%202025-03-04&q=trump%20fake%20news

sectors such as public health and the global economy. A prominent example occurred during the COVID-19 pandemic, where the spread of "infodemic" undermined public health efforts and led to real-world harm [3, 4]. Similarly, in the financial sector, the research finds different degree price impact from the fake news articles for small firms, mid-size firms, and large firms [5]. The urgency to combat fake news has never been greater, and innovative, scalable solutions are needed.

To combat this, AI has emerged as a pivotal tool, utilizing advancements in machine learning (ML) [6], deep learning (DL) [7, 8], and large language models (LLMs) [9, 10] for offering the potential to automatically identify, classify, and mitigate the spread of fake news at scale [11]. Unlike traditional manual fact-checking methods, which are time-consuming and resource-intensive, AI-driven systems can process vast amounts of data in real time, making them indispensable in the fight against misinformation [12]. The integration of AI techniques not only improves detection accuracy, but also enables the analysis of complex patterns in news content, social context, and propagation dynamics. The significance of AI-driven fake news detection thesis lies in its potential to safeguard information integrity and promote informed decision-making in the digital age. Given previous research on the influence of fake news [13, 14], by developing robust AI models, we can empower platforms and policymakers to foster a more reliable information ecosystem, protecting the public from the far-reaching consequences of digital misinformation.

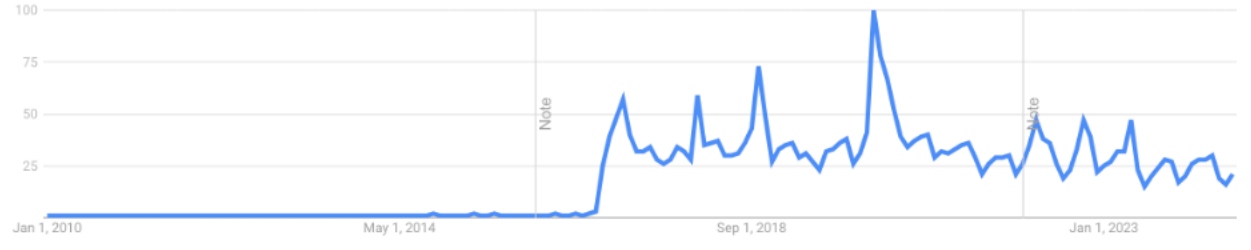

Figure 1.1: Fake News Trend 2004.2-2025.2

1.2 Related Work and Major Contributions

With the growing demand for information accuracy and integrity across social media and digital platforms, numerous studies have been conducted to improve the performance of fake news detection. Initial research on fake news detection primarily focused on psychological and sociological factors influencing misinformation dissemination. Scholars examined how cognitive biases, emotional appeal, and social influence contributed to the spread of fake news. In parallel, fact-checking platforms such as Snopes, PolitiFact, and FactCheck.org emerged to provide manual verification of news claims [15, 16, 17]. While these platforms played a crucial role in distinguishing misinformation from factual reporting, their reliance on human verification limited scalability and response speed. To address these challenges, researchers developed rule-based automated methods that analyzed linguistic patterns and stylistic features indicative of fake news [18]. However, these early approaches lacked adaptability, as misinformation continuously evolved in structure and presentation.

Later, the integration of artificial intelligence into fake news detection marked a pivotal shift, with machine learning models significantly enhancing detection capabilities. Supervised learning algorithms, such as support vector machines (SVM) and decision trees, were trained on labeled datasets to classify news articles based on textual features [19, 20]. Computational linguistics techniques, including sentiment analysis, entity recognition, and topic modeling,

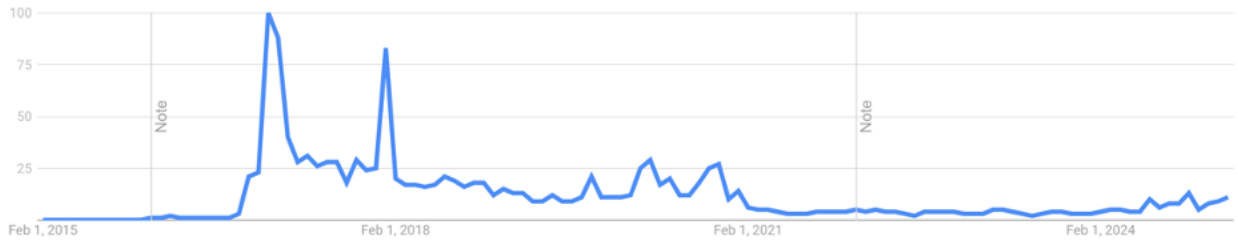

Figure 1.2: Trump Fake News Trend 2015.2-2025.2

were employed to extract meaningful insights and identify deceptive content. Despite their effectiveness, these models were constrained by the quality and diversity of training data, limiting their ability to generalize to emerging misinformation patterns. To overcome these limitations, researchers explored unsupervised learning methods, such as clustering and anomaly detection, which allowed for the identification of novel fake news stories without requiring labeled data [21]. These advancements improved adaptability to evolving misinformation tactics, enhancing the robustness of automated detection systems. However, it often lacks the granular semantic discernment required to distinguish nuance from falsehood.

To overcome these limitations, researchers have increasingly pivoted toward deep learning. The advent of deep learning further transformed fake news detection by enabling sophisticated and automated content analysis. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) facilitated the identification of manipulated images, deepfake videos, and synthetic audio, expanding the scope of misinformation detection beyond text-based approaches [22]. Transformer-based models, such as GPT-1, Bidirectional encoder representations from transformers (BERT), and T5, significantly improved the classification of misinformation by capturing complex linguistic relationships and contextual dependencies within news content [23]. These advancements enabled highly accurate detection, leveraging vast pre-trained datasets to refine AI-driven predictions.

As these architectures (like GPT, T5) scaled into Large Language Models, detection shifted from simple classification to complex reasoning, enabling the identification of logical fallacies and factual inconsistencies at a broader scale. Additionally, multimodal fake news detection emerged, integrating text, image, and video analysis to enhance detection precision and comprehensiveness [24, 25, 26]. Besides, the rise of cybersecurity, federated learning, and explainable AI (XAI) is overcoming the growing threat of "Deepfake" and manipulated media [27, 28, 29, 30]. Considering the increasing attention on fake news detection and the urgent need for innovative breakthroughs, we realize it is the right time to present a thesis about AI-driven FND.

Given the significance of news integrity, there are an increasing number of surveys reviewing FND from different perspectives. These include data mining perspectives [31, 11], fundamental theories, datasets [32], features and classification [33]. A landmark survey published in 2020 primarily reviewed traditional features, propagation-based models, and early neural network techniques [12]. As the field entered the deep learning era (2021–Present), focus shifted toward neural architectures such as CNNs, RNNs, and Transformers (e.g., BERT) to capture hierarchical and contextual features [22]. Besides, a survey about multimodal fake news detection showed it effectively combines image analysis and text processing, exhibited a superior performance [34].

Despite surveys mentioned the classification of fake news detection and mitigation [33], or focused on a specific topic like the use of deep learning methods for multimodal fake news detection [35, 36], they all make a contribution to the development of this area. Existing surveys lack a comprehensive synthesis of the "full pipeline" specifically tailored to the era of LLMs. Furthermore, previous literature has seldom maintained a strict distinction between the specific utilization of approaches and techniques. To bridge this gap, this thesis introduces a distinction: **FND Approaches (Model-centric):** These represent the overarching classification frameworks (e.g., Traditional Approaches, Machine Learning, Deep Learning, and Trending Approaches). **FND Techniques (Process-centric):** These refer to the operational stages within the detection pipeline (e.g., Dataset, Data Augmentation, Information Extraction, Model Validation and Refinement, and Results Explanation and Feedback).

Therefore, by categorizing existing methodologies in this way, we offer a more precise understanding of how structural models and procedural optimizations interact to fight fake news. This thesis benefits researchers and practitioners who want to keep up with the state-of-the-art research in AI-driven FND. The major contributions of this thesis are as follows:

- To provide a holistic conceptual framework, we synthesize the fake news ecosystem by clarifying its definition, related terms and distinctive characteristics.

- To trace the historical trajectory of fake news detection approaches, we review the evolution of fake news detection, spanning from early traditional approaches to advanced machine learning and deep learning, eventually culminating in contemporary LLM-based reasoning stages.

- To categorize and evaluate AI-driven fake news detection approaches, we present a comparison of models across different modalities. This includes traditional machine learning and deep learning, highlighting their respective advantages, limitations, and applications in fake news detection.

- An evaluation of AI-driven fake news detection techniques is provided, which includes key stages of the detection process: dataset, data augmentation, information extraction, model validation and refinement, and results explanation and feedback. Moreover, we introduce a structured framework to illustrate the pipeline work.

- To complement the theoretical analysis, a focused empirical study is provided, which utilizes the BuzzFeed News dataset from the 2016 U.S. presidential election. It employs rigorous Exploratory Data Analysis (EDA) to delineate the distinct linguistic patterns and structural features that differentiate fake narratives from authentic reporting. Furthermore, we systematically evaluates the performance of various machine learning classifiers, assessing their efficacy in veracity detection.

- We discuss the main challenges and trends in fake news detection from four perspectives. By highlighting these challenges, we aim to guide future research towards overcoming these obstacles and advancing the field.

1.3 Method of Paper Collection

The thesis focuses on reviewing fake news detection from the perspective of AI-driven approaches (model-centric) and AI-driven techniques (process-centric), so we retrieved top journals such as ACM Trans. Inf. Syst., ACM Trans. Knowl. Discov. Data, IEEE Trans. Knowl. Data Eng, etc., and top conferences such as SIGIR, KDD, WWW, RecSys, WSDM and so on. Leveraging scholarly databases like dblp and Google Scholar, we systematically conducted searches employing specific keywords like "fake news", "misinformation", "fake news detection", "misinformation detection" to search related work. In order to make the retrieved papers more relevant, we also used keywords that related to AI, such as "machine learning", "deep learning", and "Large Language Models" to get more papers.

After the initial search, we carefully curated a selection of relevant papers for further examination. Included papers were peer-reviewed English-language articles that addressed the approaches or techniques of fake news detection. These articles, authored by academic researchers, were recent. Excluded were articles in languages other than English, those not focused on the mentioned topics, and articles that did not adhere to these inclusion criteria. Then, based on the above retrieved papers related to fake news detection, we illustrate the statistics of them according to the publication time and papers classification, as shown in

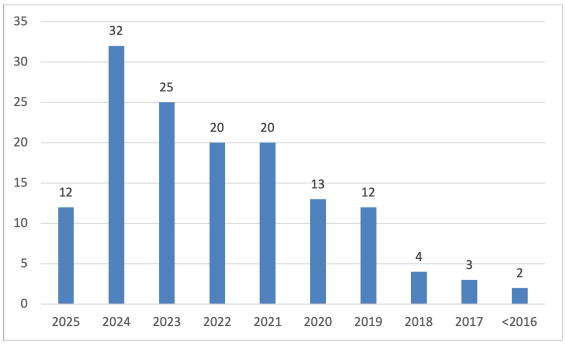

Figure 1.3: The Statistics of Papers with the Publication Year

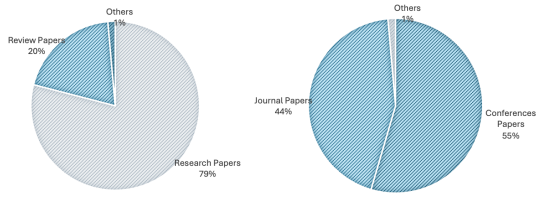

Figure 1.4: The Statistics of Papers with Classification

Figure 1.3 and Figure 1.4 respectively.

1.4 Thesis Organization

The remainder of this thesis is organized as follows, Figure 1.5 presents the main parts of this thesis. By structuring the thesis in this manner, we aim to provide a holistic understanding of AI-driven fake news detection, providing valuable insights to researchers, practitioners, and policymakers alike.

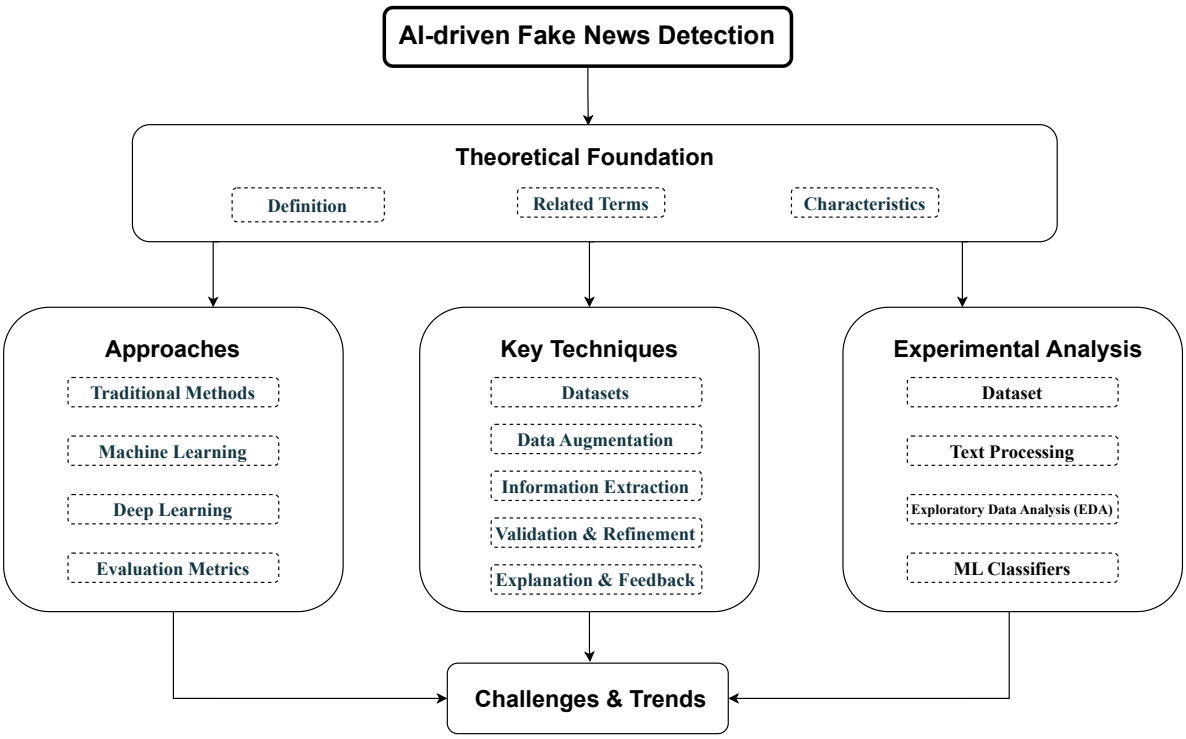

Figure 1.5: Main Parts of the Thesis

chapter 2: Theoretical Foundations of Fake News Detection This chapter delves into the theoretical foundations of fake news detection, including the definition of fake news, related terms, and the characteristics of fake news.

chapter 3: AI-driven Approaches for FND This chapter provides a detailed analysis of AI-driven approaches, covering traditional approaches, machine learning, deep learning, and trending approaches.

chapter 4: AI-driven Techniques for FND This chapter examines the key techniques: dataset, data augmentation, information extraction, model validation and refinement, and results explanation and feedback.

chapter 5: Experimental Results and Analysis This chapter outlines the methodology on Machine Learning-Based Fake News Detection. It combines meticulous data collection, exploratory data analysis, and the application of various machine learning classifiers. The methodology is structured to not only identify linguistic patterns and features unique to fake news but also to compare the accuracy of fake news detection.

chapter 6: Challenges, Future Trends and Conclusion This chapter discusses some challenges and future trends from 4 aspects: (1) data evolution and quality, (2) model interpretability and ethical concerns, (3) multimodal complexity, (4) scalability, adaptability and others. Then, this chapter concludes the thesis and discusses the future work.

Chapter 2

Theoretical Foundations of Fake News Detection

In this chapter, we introduce the definition of fake news, compare related terms with fake news, map the fake news ecosystem, and examine its multi-dimensional characteristics.

2.1 Fake News Definition

The definition of fake news has evolved over time, reflecting its diverse manifestations across topics, styles, and platforms [11]. In 2016, a group of Stanford researchers, concerned about the spread of fake news during the Trump election, defined fake news as: news articles that are intentionally and verifiably false and could mislead readers [31], and the study noted that the dissemination of fake news has had a significant impact on politics. Subsequently, other scholars have also studied fake news in the fields of economics and health, and proposed different definitions. For example, more emphasis has been placed on the deliberate and deceptive nature of fake news, which some scholars define as any false information or story

posted on the Internet for the purpose of misleading readers [37]. According to Wikipedia[3], fake news, or information disorder, encompasses false or misleading information, including misinformation, disinformation, propaganda, and hoaxes, presented as news. In this thesis, we define fake news as any information or statement that is inconsistent with its factual content, encompassing a wide range of concepts such as deceptive journalism, fake news, false information, misinformation, clickbait, and rumors.

With reference to Wikipedia, we map the fake news ecosystem based on the entire process of fake news generation, dissemination and use, as shown in Figure 2.1.

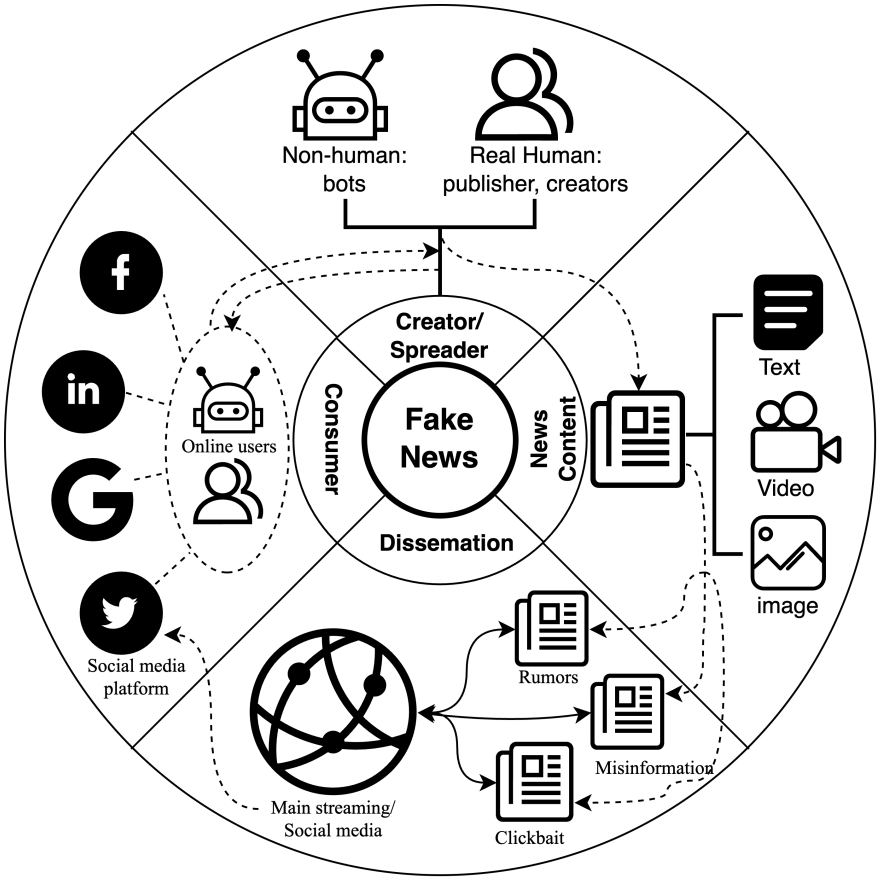

Figure 2.1: Fake News Ecosystem

[3]https://en.wikipedia.org/wiki/Fake_news?useskin=vector

The fake news ecosystem is a complex, multi-layered environment where various actors and platforms interact to produce and propagate misinformation. This cycle begins with creators and spreaders, which include both non-human agents (automated bots) and real humans, such as motivated publishers or unwitting users. These entities generate news content in diverse formats—ranging from traditional text to sophisticated deepfake videos and manipulated images. Once produced, this content is introduced to consumers via mainstream and social media platforms. The dissemination phase is characterized by a feedback loop where user engagement—such as shares, comments, and likes—further amplifies the reach of deceptive narratives. This ecosystem is not static; it is a dynamic process where misinformation, rumors, and clickbait are continuously recycled and reshaped, making detection a moving target that requires adaptive AI-driven solutions.

2.2 Related Terms

In addition to fake news, related terms are often used interchangeably. These terms collectively highlight the spectrum of false information, ranging from unintentional errors to malicious fabrications.

Deceptive News refers to news that is deliberately fabricated and demonstrably false. False news consists of reports disseminated through media that contain certain non-factual elements. Satirical news typically incorporates satirical and exaggerated reporting on current events, often including exaggerated or false content. Misinformation refers to information that has been intentionally or unintentionally distorted or entirely fabricated with the aim of deceiving and misleading the public. Clickbait usually consists of sensationalized headlines designed to attract users' attention and drive them to click on a webpage, generating advertising revenue through user engagement. Rumors are often emotionally charged personal narratives, and when proven false, they constitute fake news. Besides, false information and elective acceptance are related terms of fake news.

Concepts	Authenticity	Intent	Degree of Harm
Deceptive news	Not true	Misinformation	Intentionally written, high impact
False news	Not true	Misinformation	Somewhat influential
Satirical news	Not uniform	Entertaining the public	Less influential
False information	Not true	Misleading the public	Maliciously disseminated, affecting society and the public
Misinformation	Not true	Misinformation	Somewhat influential
Selective Acceptance	Usually true Facts	Misleading the public	Biased absorption by users, affecting users personally
Clickbait	Usually not true	Profit from user clicks	Headline party, average impact
Rumor	Not true	Attract user attention, mislead the public	Large, affecting the physical and mental health of the public

Table 2.1: Terms and Concepts Related to Fake News

Based on these definitions and conceptual distinctions, fake news can be analyzed from three dimensions: authenticity, intent, and degree of harm. Authenticity pertains to whether the content contains any non-factual statements. Intent refers to the motives of users who publish or share fake news, such as misleading or entertaining the public or seeking financial gain. The degree of harm measures the extent to which fake news negatively impacts society or the general public. Table 2.1 provides a summary of these concepts.

2.3 Fake News Characteristics

Fake news exhibits distinct characteristics that differentiate it from genuine news, we can conclude them into three primary dimensions: user characteristics, content characteristics, and dissemination characteristics [12].

2.3.1 User Characteristics

User characteristics refer to the attributes and behaviors of individuals who create and disseminate news on social media platforms. These characteristics can be analyzed to enhance the accuracy of fake news detection models.

Attribute characteristics include user profile details such as verification status, follower count, number of accounts followed, completeness of personal information, account registration date, educational background, profession, and age. Research suggests that users with lower educational attainment are more likely to create and propagate fake news, while older users are more susceptible to believing and further disseminating it. Additionally, fake news creators often use newly registered and disposable accounts to spread misinformation [38].

Behavioral characteristics pertain to user interactions with news, including commenting, sharing, and engagement patterns after publication. Notably, individuals spreading fake news exhibit behavioral anomalies, such as a high daily posting frequency and irregular retweeting intervals [12]. These behaviors can serve as indicators for detecting fake news.

Political bias significantly influences user-profiles and their choices in news consumption. Studies in media sociology have demonstrated correlations between partisan bias and perceptions of news authenticity. Political affiliation is closely related to both user attributes and behavior, and incorporating political stance as an auxiliary feature can enhance fake news detection [39].

2.3.2 Content Characteristics

Content characteristics of fake news can be categorized into emotional, textual and semantic, and multimodal characteristics.

Emotional Characteristics: One of the most prominent characteristics of fake news is its emotional tendency, often employing emotionally charged language to evoke strong reactions such as fear, anger, or excitement [40]. This emotional manipulation is designed

to capture the reader's attention and increase the likelihood of sharing, thereby amplifying its reach. Unlike factual news, which maintains a neutral stance, fake news reflects personal opinions and strong emotional polarity, frequently using exaggerated and subjective language. Dickerson [41] found that emotion-related behaviors can distinguish between human accounts and social bot accounts. For example, real news reports tend to be neutral and devoid of personal sentiment, while fake news often consists of social media users expressing personal opinions on an issue, frequently containing highly polarized emotional terms and strong subjective tones.

Textual and Semantic Characteristics: Fake news is characterized by linguistic styles that include sensational headlines, exaggerated claims, and informal language. These stylistic elements are intentionally crafted to make content appear more engaging and credible, even when the underlying information is false or misleading. Additionally, fake news often contains semantic inconsistencies, such as logical contradictions, factual inaccuracies, and a lack of credible sources.

Rubin [42] analyzed the rhetorical structures, discourse components, and coherence relationships of both real and fake news, applying a vector space model to cluster news based on discourse feature similarity. Their research found that fake news tends to avoid stating facts directly and instead relies heavily on exclamation marks, question marks, adjectives, and first-person pronouns, making the narrative less logical and more emotionally appealing. Similarly, Brasoveanu [43] extracted sentiment, entity, and fact-based relationships from news texts, using semantic and syntactic feature analysis to identify fake news. Wang [44] further refined FND by capturing event theme sentences, statements that reflect factual elements of an event, and summarizing them using a dual similarity approach. They identified four fine-grained discrepancies in sentence definitions corresponding to common deceptive strategies in fake news, designing quantitative methods to measure these discrepancies. By converting these differences into sentence weights and integrating them into a BERT pre-trained model, they enhanced the accuracy of fake news representation and detection.

Multimodal Characteristics: Fake news increasingly incorporates multimodal features, such as manipulated images, videos, or audio, to enhance its perceived credibility and emotional impact. The use of multimedia elements not only makes the content more persuasive but also complicates the detection process, as it requires the analysis of multiple data types. Deepfake techniques, such as face-swapping and voice simulation, have made fake news more deceptive than static text-based misinformation [45]. However, due to the challenges associated with extracting meaningful features from large-scale multimedia data, research on multimodal fake news detection remains limited. Jing [46] proposed a progressive fusion network for multimodal fake news detection, capturing text and image features at different levels and enhancing cross-modal relationships, thereby improving detection performance.

2.3.3 Dissemination Characteristics

Dissemination characteristics refer to the patterns of information propagation, including transmission paths, speed, and reach.

The propagation of fake news is heavily influenced by its dissemination dynamics, which are shaped by the structure and behavior of social networks. Fake news often spreads through propagation paths that leverage user interactions such as shares, likes, and comments. These interactions create a network effect, where the content is rapidly disseminated across platforms, reaching a wide audience in a short period. The propagation speed of fake news is another critical factor, as it tends to spread faster than genuine news due to its sensational nature and emotional appeal [47]. This rapid dissemination is further amplified by algorithms on social media platforms that prioritize engaging content, often at the expense of accuracy.

The fake news lifecycle [48] follows a distinct trajectory: creation, propagation, and impact. Research has revealed that a substantial portion of rumor transmission occurs before verification, often experiencing viral growth within minutes of its initial release. As verification efforts increase, fake news loses traction, highlighting the importance of early

detection mechanisms.

To better understand the spread of fake news, researchers have developed various propagation models that analyze its dissemination patterns. One such model is the epidemiological model [49], which treats fake news as a contagion and examines its spread through networks based on user interactions. This model highlights the role of influential users and network structures in amplifying fake news, similar to how diseases spread through populations. Another approach is the information cascade model, which focuses on the sequential spread of information and identifies key nodes in the network that drive the propagation process [12]. These models provide valuable insights into the mechanisms underlying the spread of fake news and offer a foundation for developing strategies to mitigate its impact.

Additionally, the propagation scope of fake news is expanded by the presence of echo chambers and filter bubbles, where users are exposed primarily to information that aligns with their existing beliefs [50]. These mechanisms create an environment where fake news can thrive, as users are less likely to encounter contradictory information that could challenge its validity.

A thorough understanding of these features facilitates the development of more targeted fake news detection models, thus mitigating their adverse effects.

Chapter 3

AI-Driven Approaches for FND

In this chapter, we mainly discuss and compare FND approaches (model-centric), which represent the overarching classification frameworks—including Traditional Approaches, Machine Learning, Deep Learning, and Multimodal Approaches—that serve as the primary structural paradigms for fake news detection, as shown in Figure 3.1. This section evaluates these paradigms, critically analyzing their specific mechanisms, strengths, and inherent limitations in addressing fake news detection.

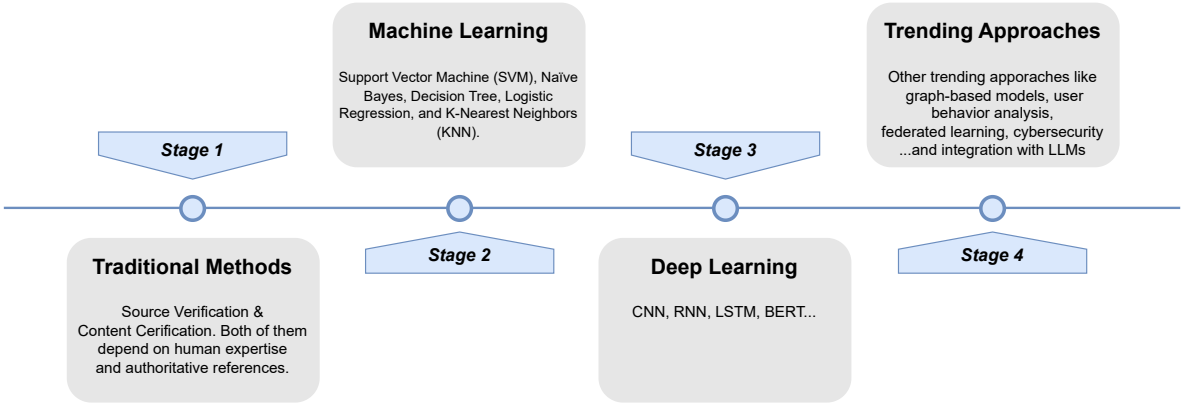

Figure 3.1: The Evolution of Fake News Detection Approaches

3.1 Traditional Methods

Traditionally, the identification of false information relied on manual processes conducted by experts and established organizations. These methods center on human judgment and the use of authoritative references to confirm the truth of a claim. We can divide these traditional strategies into two main categories, which are source verification and content verification. Both categories depend on human expertise and the use of trusted records to validate news [51].

Figure 3.2: The Screenshot of Factcheck.Org

Source verification focuses on the reliability of the news outlet. Fact-checkers evaluate the history and reputation of a source to determine if it consistently produces accurate information. Professional journalistic standards and established news organizations act as the standard for this evaluation. While this method helps identify known biased actors, it

often fails when dealing with misinformation from newly formed platforms or less regulated online spaces where the history of the publisher is unknown. Content verification involves

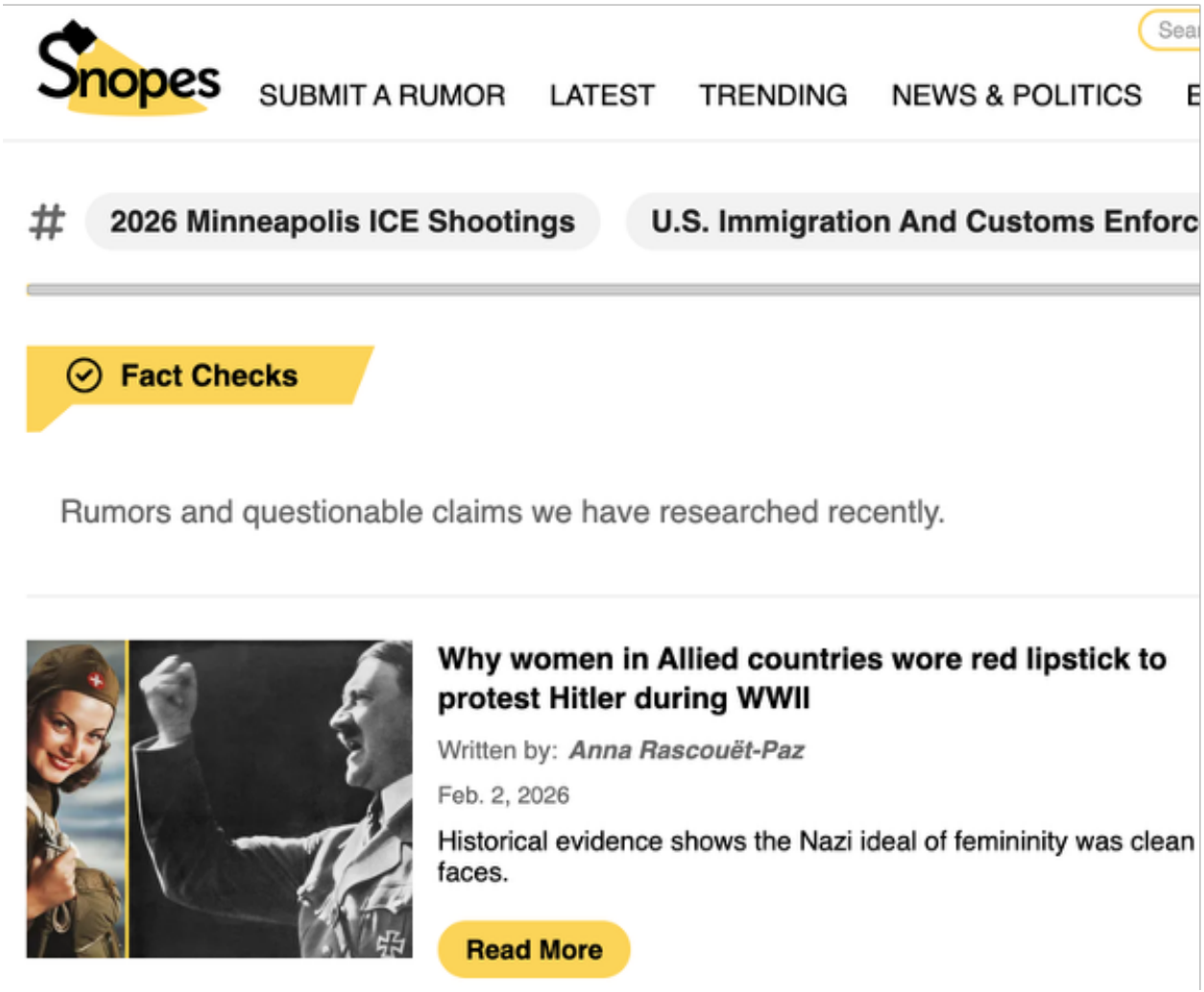

Figure 3.3: The Screenshot of Snops.com

the direct examination of the information. Experts in specific areas like medicine or history use their knowledge to check the accuracy of claims. Traditional organizations also compare news content against established databases of known facts. Well-known platforms such

as FactCheck.org [4] and Snopes.com [5] represent this approach. FactCheck.org serves as a nonprofit organization that monitors the accuracy of political statements in the United States. It uses expert analysis to evaluate claims from debates and social media while also managing specialized branches like SciCheck for scientific claims. Similarly, Snopes.com addresses a broad range of topics from urban legends to business fraud. It uses professional knowledge to assign truth ratings to printed and digital resources. These platforms show the value of human-led verification in maintaining information integrity. Figure 3.2 and Figure 3.3 show the content of websites.

Although these manual methods provide high accuracy, they face significant limitations in the current information environment. The manual process is slow and requires a large amount of human labor, which makes it difficult to scale. The high financial cost of hiring experts further restricts the ability of these organizations to monitor the large volume of news generated every day [37]. Most importantly, the speed at which fake news spreads online exceeds the capacity of human fact-checkers. Because the nature of fake news changes rapidly, relying only on manual verification is insufficient for real-time detection [51]. These difficulties suggest that while manual verification provides a strong foundation, there is a clear need for automated systems to improve the speed and reach of fake news detection.

3.2 Machine Learning

Machine learning has emerged as a powerful tool for detecting fake news since 2017 [52]. Various ML algorithms, including Naïve Bayes, Support Vector Machine, Decision Tree, Logistic Regression and K-Nearest Neighbors (KNN), have been extensively applied in this domain, each with distinct strengths and limitations. To provide a comprehensive comparison of these approaches, Table 3.1 summarizes their methodologies, advantages, limitations, and

[4] https://www.factcheck.org/
[5] https://www.snopes.com/

Model	Strengths	Weakness	Application
Naïve Bayes [20, 52]	Fast to train and perform inference, requires less data	Reduce accuracy in complex datasets	Spam filtering, sentiment analysis
SVM [20, 53]	Effective in high-dimensional spaces, robust to overfitting	Computationally expensive for large datasets, sensitive to the choice of kernel and hyperparameters.	Text classification, image recognition, bioinformatics
Decision Tree [54, 55]	Intuitive and interpretable, handles both numerical and categorical data	Prone to overfitting, sensitive to noisy data, and may generate complex trees that lack generalization	Credit scoring, fraud detection, medical decision-making.
Logistic Regression [54, 56]	Simple, interpretable, computationally efficient, works well with linearly separable data.	Limited in handling complex relationships, assumes linearity in feature-space interactions	Binary classification tasks
K-Nearest Neighbors [57, 55]	Capture complex decision boundaries	High computational costs for large datasets, negatively impact classification accuracy	Clustering patterns

Table 3.1: Comparative Analysis of Machine Learning Model-based FND

applications in fake news detection.

3.2.1 Naïve Bayes

Naïve Bayes is a probabilistic classifier based on Bayes' theorem, which assumes feature independence. This simplification allows for efficient and scalable classification, making it particularly useful for high-dimensional text classification tasks such as spam detection and sentiment analysis. Researches [20, 52] highlight its application in fake news detection, showing its ability to process vast amounts of textual data efficiently (accuracy: 75%). However, the assumption of feature independence is often unrealistic, potentially leading to suboptimal classification performance when features exhibit strong dependencies.

3.2.2 Support Vector Machine

SVM is a supervised learning algorithm primarily used for classification tasks. It constructs an optimal hyperplane that maximizes the margin between different classes, enabling robust classification even in complex and high-dimensional spaces. By utilizing different kernel functions such as polynomial or radial basis function (RBF), SVM can handle both linear and non-linear classification problems. Studies [20, 53] have demonstrated the effectiveness of SVM in fake news detection, particularly in scenarios where textual data exhibit clear separable patterns. However, SVM can be computationally expensive for large datasets and may struggle with noisy or overlapping classes.

3.2.3 Decision Tree

Decision Tree algorithms operate by recursively splitting the dataset based on feature values, forming a hierarchical structure that facilitates interpretability. Each node represents a decision rule, and leaves correspond to final classifications. Decision Trees are advantageous due to their transparency and ease of implementation. Krishna had applied Decision Trees to fake news detection, demonstrating their ability to capture nonlinear patterns [54, 55]. Nonetheless, Decision Trees are prone to overfitting, particularly with deep trees, which can reduce generalizability when applied to new data.

3.2.4 Logistic Regression

Logistic Regression is a statistical method used for binary classification, predicting the probability of a sample belonging to a particular class. By applying the logistic function, it transforms continuous values into probabilities, allowing for threshold-based classification. The algorithm's simplicity and interpretability have made it a widely used baseline model in fake news detection. Krishna also showed its effectiveness, particularly in well-structured datasets where linear separability holds [54, 56]. However, Logistic Regression may struggle

with complex relationships and non-linearly separable data, limiting its efficacy in nuanced fake news detection tasks.

3.2.5 K-Nearest Neighbors

KNN is a non-parametric algorithm that classifies data points based on their proximity to labeled instances in the feature space. It is particularly effective when the decision boundary is highly nonlinear, making it useful for detecting fake news by identifying similarities in textual and structural features [57, 55]. However, KNN suffers from high computational costs as dataset size increases, and it is sensitive to noisy or irrelevant features, which can impact classification performance.

3.3 Deep Learning

Deep learning is a subfield of machine learning strategies, which displays high precision and exactness in fake news detection. Unlike traditional machine learning, deep learning models can automatically extract hierarchical features and learn contextual relationships, making them particularly effective for misinformation detection. For example, convolutional neural networks, recurrent neural networks, Long Short-Term Memory (LSTM), Bidirectional Encoder Representations from Transformers are broadly utilized ideal models for deep learning based FND. Additionally, trending approaches such as graph-based models, user behavior analysis, cybersecurity integration, federated learning and other LLMs-based models further enhance the ability to detect fake news. This section outlines these models' advantages and limitations, and their application in recent research. We summarize them in Table 3.2.

Model	Methodology	Strengths	Weakness	Application
CNNs [58, 59, 60, 61]	Image and Video Analysis	Effective at capturing spatial hierarchies, translation invariance, and feature extraction.	Requires large labeled datasets, struggles with sequential data, and is computationally expensive.	Image recognition, object detection.
RNNs [60, 62, 63]	Sequential Data Analysis	Effective for modeling sequential dependencies, suitable for time-series and language tasks.	Prone to vanishing/exploding gradient problems, limited in capturing long-term dependencies.	Speech recognition, language modeling, machine translation.
LSTM [58, 60]	Sequential Data Analysis	Handles long-term dependencies, effective for sequential data with complex temporal relationships.	Computationally intensive, requires more training time, and is less interpretable.	Text generation, sentiment analysis, machine translation, speech recognition.
BERT and Transformer Based [64, 65]	Contextual Language Understanding	Captures deep contextual relationships, improves performance on various NLP tasks, and is transferable to different domains.	Requires significant computational resources, complex to fine-tune, and prone to overfitting on small datasets.	sentiment analysis, named entity recognition, text summarization.

Table 3.2: Comparative Analysis of Deep Learning Model-based FND

3.3.1 Convolutional Neural Networks

CNNs are deep learning architectures primarily designed for structured grid data, such as images. However, it has also been successfully adapted for text classification tasks, including fake news detection. CNNs process text by representing it as matrices and applying convolutional filters to extract key features. The hierarchical learning of spatial and contextual patterns enables CNNs to detect fake news efficiently. Studies [58, 59, 60, 61] have applied CNNs in misinformation classification, demonstrating their ability to capture local dependencies in textual data. Despite their efficiency, CNNs have limitations in capturing long-range dependencies and sequential relationships in text, making them less effective for tasks requiring deep contextual understanding.

3.3.2 Recurrent Neural Networks

RNNs are specifically designed for sequential data, making them highly suitable for analyzing textual information in fake news detection. They incorporate a memory mechanism that allows information from previous time steps to influence current predictions, enabling the network to model context effectively. Research [60, 62, 63] has shown that RNNs perform well in detecting fake news by analyzing sentence structures and contextual cues. However, traditional RNNs suffer from issues like vanishing and exploding gradients, which can hinder their ability to capture long-term dependencies in lengthy texts.

3.3.3 Long Short-Term Memory

LSTM is a specialized type of RNN designed to address the limitations of standard RNNs by incorporating memory cells and gating mechanisms (input, forget, and output gates). These gates control the flow of information, allowing LSTM networks to retain or discard data over long sequences. This capability makes LSTMs particularly effective for fake news detection, as they can analyze long textual sequences while preserving contextual meaning. Studies [58,

60] have highlighted LSTM's superior performance in handling misinformation by leveraging its ability to learn from past textual data. However, LSTMs require significant computational resources and can be slower to train compared to other deep learning models.

3.3.4 BERT

BERT is a transformer-based model that significantly enhances deep learning applications in natural language processing (NLP). Unlike RNNs and LSTMs, which process text sequentially, BERT employs a bidirectional approach, considering both preceding and succeeding words to capture richer contextual representations. This attribute makes BERT highly effective in fake news detection, as it can understand nuanced meanings and contextual relationships in text. Studies [64, 65] have successfully utilized BERT for misinformation classification, demonstrating improved accuracy over traditional NLP methods. However, BERT's high computational cost and requirement for extensive pretraining make it resource-intensive.

3.4 Trending Approaches based on DL

With the continuous advancement of deep learning, several emerging approaches have been developed to enhance the effectiveness and robustness of fake news detection. These methods go beyond traditional deep learning architectures and integrate novel approaches to address the complexities of misinformation propagation. This section discusses key trending approaches, including graph-based models, user behavior analysis, cybersecurity integration, federated learning, explainable AI, and multimodal analysis. To provide a comprehensive comparison of these approaches, Table 3.3 summarizes their methodologies, advantages and disadvantages.

Graph-based models have gained significant attention in fake news detection due to their ability to represent and analyze relationships among news articles, social media users, and shared content. These models leverage graph structures where nodes represent entities

Technique	Methodology	Strengths	Weakness
Graph-based Models [66, 67]	Representation of relationships in graphs	Effectively models complex relationships, preserves structural information, and generalizes well to networked data.	Computationally intensive for large graphs, requires domain-specific feature engineering, and may suffer from over-smoothing.
User Behavior Analysis [68, 69]	Studying engagement patterns	Enhances personalization, improves recommendation accuracy, and detects anomalies in user activity.	Privacy concerns, potential biases in behavioral data, and challenges in handling evolving user behaviors.
Cybersecurity Integration [27]	Adapting techniques against adversaries	Improves threat detection accuracy, enables real-time security monitoring, and reduces manual intervention.	High false positive rates, adversarial attacks on AI models, and computational resource constraints.
Federated Learning [28, 70]	Collaborative learning across platforms	Enhances data privacy, reduces the need for centralized data storage, and enables collaborative learning.	High communication overhead, challenges in model synchronization, and vulnerability to adversarial attacks.

Table 3.3: Comparative Analysis of DL Trending Model-based FND

(e.g., news articles or users), and edges represent interactions (e.g., content sharing or user engagement). By applying Graph Neural Networks (GNNs), these models can learn structural patterns, propagation process and detect misinformation based on network dynamics [67]. Mayank [66] showed an approach that combines natural language processing and tensor decomposition model to encode news content and embed Knowledge Graph (KG) entities, respectively. DEAP-FAKED obtains an F1-score of 88% and 78% for the two datasets, which is an improvement of 21%, and 3%, respectively, which shows the effectiveness of the approach. However, these models require high-quality relationship data and may struggle with evolving misinformation strategies that manipulate network structures.

Fake news detection can benefit significantly from analyzing user behaviors on digital platforms. Karine presents a user-centric theoretical model that elucidates the factors that allow online users to identify fake news within the social [68]. Anothter research proposed a framework named UPFD, which simultaneously captures various signals from user preferences by joint content and graph modeling. Experimental results on real-world datasets demonstratethe effectiveness of the proposed framework [69]. However, privacy concerns and evolving user behaviors pose challenges to this approach, requiring continuous adaptation of detection strategies. Cybersecurity integration enhances fake news detection by incorporating cybersecurity frameworks, such as anomaly detection and threat intelligence, to identify coordinated misinformation attacks [27]. Meanwhile, federated learning offers a privacy-preserving approach by enabling collaborative model training across decentralized data sources without sharing sensitive user information [28, 70]. These methods contribute to a more robust and privacy-conscious fake news detection ecosystem but face challenges related to scalability and security risks in distributed environments.

Recent advancements extend beyond traditional deep learning models by integrating Large Language Models with other trending approaches. One significant direction involves coupling LLMs with graph-based models to analyze relationships and propagation structures within social networks [67]. Systems also combine LLMs with user behavior analysis to

identify patterns in how individuals interact with and share information [69]. Cybersecurity protocols are increasingly embedded within these frameworks to protect the detection models against adversarial manipulation. Federated learning is paired with LLMs to address privacy restrictions while training on distributed data without centralizing sensitive information [71]. Additionally, explainable AI components are attached to LLM outputs to interpret the reasoning behind classification results and improve trust. Multimodal strategies link LLMs with image and video processing tools to detect inconsistencies across different data types. These hybrid approaches lead into the detailed discussion of Large Language Model based detection methods. We will discuss explainable AI and multimodal from LLMs-based techniques angle (As shown in Chapter 4.2.3 and Chapter 4.5) and challenge aspects (As shown in Chapter 6.2 and Chapter 6.3) in the following thesis.

3.5 Evaluation Metrics

Following the construction of a fake news detection models, the evaluation process becomes a critical step in assessing the model's effectiveness. While a model may achieve high classification accuracy after training, accuracy alone is insufficient to determine its applicability across different contexts. To ensure a comprehensive assessment, researchers typically employ multiple evaluation metrics and construct a confusion matrix, which provides an intuitive representation of the model's overall performance on the test set, including true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Among the various evaluation metrics, the F1-score is the most commonly used indicator for assessing the performance of automated claim verification systems. Additionally, other metrics such as precision (P), recall (R), accuracy (A) and ROC are also employed to evaluate system performance, as shown in Table 3.4.

Beyond conventional performance metrics, XAI have increasingly been integrated into the evaluation of fake news detection models. Methods such as G-Eval [77], Local Interpretable

Metric	Formula	Description	Role in Fake News Detection
Recall [72, 73]	$\frac{TP}{TP+FN}$	Proportion of actual fake news correctly identified	Measures the ability to find all misinformation
Precision [72]	$\frac{TP}{TP+FP}$	Proportion of predicted fake news that is truly fake	Measures the reliability of flagging suspicious content
F1 Score [73, 74]	$2 \times \frac{Precision \times Recall}{Precision+Recall}$	Harmonic mean of precision and recall	Provides a single score for overall classification quality
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Ratio of correct predictions to total instances	Offers a general overview of model performance
ROC AUC [75, 76]	$\int_0^1 TPR(FPR^{-1}(t))dt$	Area under the receiver operating characteristic curve	Assesses discrimination capability across thresholds

Table 3.4: Evaluation Metrics of FND Models

Model-agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), and Integrated Gradients facilitate interpretability by quantifying feature importance and assessing rule precision and coverage [78, 65]. The selection of these evaluation methods should be tailored to the specific characteristics and implementation strategies of the framework, ensuring the reliability and adaptability of the fake news detection system.

Chapter 4

AI-Driven Techniques of FND

This chapter turns to the operational stages within the detection pipeline, which we define as fake news detection techniques. In this thesis, techniques represent the process-centric operational stages such as dataset, data augmentation, information extraction, model validation and refinement, and results explanation and feedback. These stages are important because they determine how models handle raw information and transform it into reliable predictions. Figure 4.1 shows a structured framework that illustrates how these techniques connect to form a complete detection system. The following subsections examine each stage in detail.

4.1 Datasets

In the framework of LLM-based fake news detection, the datasets employed by researchers may vary depending on factors such as data collection platforms, content types, and whether propagation metadata is recorded.

However, there are two main challenges, one is the absence of comprehensive benchmark datasets with reliable ground-truth labels, and the scarcity of sufficiently large datasets capable of supporting robust analytical procedures has been identified as a primary constraint [22].

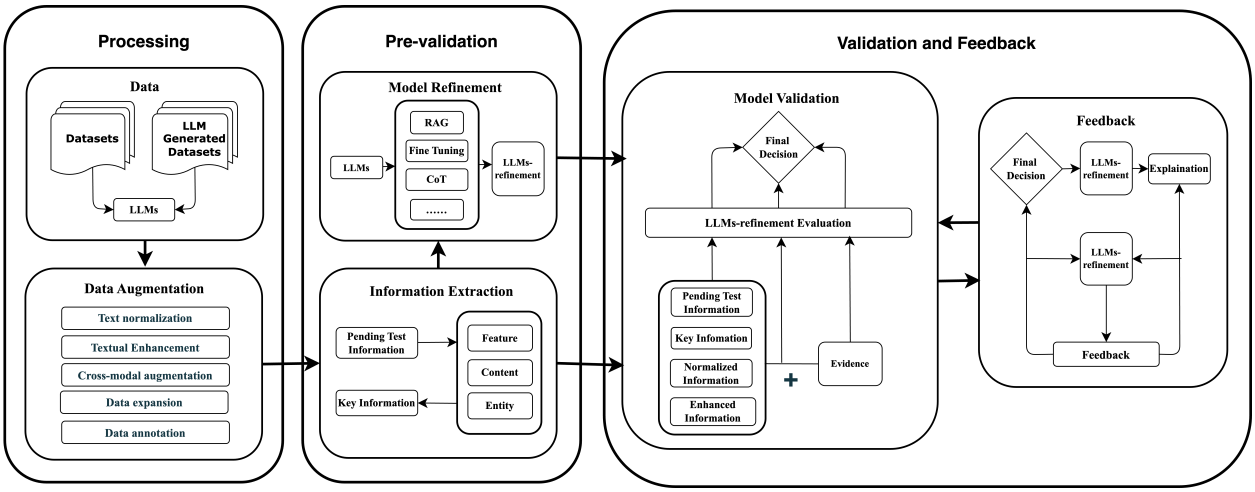

Figure 4.1: The Framework Diagram of LLM-based FND Techniques

Conventional fake news detection models typically rely on pre-collected datasets containing both genuine and fabricated news items for training purposes. Nevertheless, these datasets are often limited in scale and fail to encompass the full diversity and complexity of misinformation. Furthermore, the processes of manual annotation and misinformation collection are both time-intensive and costly.

To address these limitations, researchers have begun exploring the use of LLM-generated synthetic data as a potential solution. This section provides a overview of current data sources, selection criteria, and input pipelines within LLM-based fake news detection frameworks, with particular emphasis on LLM-generated datasets.

4.1.1 Commonly Used Datasets

The data sets used to detect fake news on social networks originate primarily from four main platforms: Twitter, Facebook, Reddit, and Weibo. In the field of fake news detection, more than half of the datasets employ three veracity labels: true, false, and unverified, while others adopt a binary classification (true vs. false) [79].

Dataset	Type	Source	Size	Modality	Introduction & Characteristics
PHEME [87]	Rumor, Non-rumer	Twitter	6425	Text	Integrates tasks such as rumor detection, position classification, and accuracy prediction to support multi-task learning for rumor validation.
Twitter15 [81]	Social Media Data	Twitter	1490	Propagation Tree	Distinguishes between rumors and non-rumors, and also subdivided into true rumors, false rumors, and unconfirmed rumors, facilitating more accurate identification of rumor types.
Twitter16 [82]	Social Media Data	Twitter	818	Propagation Trees	Supports early detection of rumors, incorporating text features, user features, and communication structure features to provide multi-dimensional information for rumor detection.
BuzzFeedNews [88]	Social Media Data	Facebook	2282	Text	Includes political information
FakeNewsNet [83]	Fake News	Twitter	23196	Text	Provides bot ratings and user interaction data.
FEVER [80]	Claims	Wikipedia	185,445	Text	Allows complex multi-hop reasoning
LIAR [89]	Claims	POLITIFACT	12.8k	Text	Contains a broad political statement.

Table 4.1: Commonly Used Datasets

In LLM-based fake news detection frameworks, researchers frequently utilize datasets such as: FEVER [80], Twitter15 [81] and Twitter16 [82], GossipCop and PolitiFact [83], LIAR [84], SciFact [85], and Snopes [86]. These resources enable the systematic evaluation of detection techniques across various textual formats. (A detailed comparison of commonly used datasets is provided in Table 4.1.)

As social media content grows increasingly complex, researchers have begun incorporating non-traditional data sources. For instance: DAT@Z21 [90] supports multimodal fake news detection by providing a diverse set of features, including textual content, social context, user engagement, spatiotemporal metadata, and visual elements. The Long-Text Chinese Rumor (LTCR) dataset [91] addresses challenges in long-form misinformation. It is particularly effective for detecting multidimensional fake news, especially in global issues like COVID-

19. PolitiFact and similar datasets specialize in political discourse analysis. The CLAN dataset [92] offers a real-world social media corpus with 6,000+ posts, each annotated with standardized claims.

Despite these datasets are from real world, they have limitations in data quality (We will disscuss it in Section 6.1). Thus, LLM-generated synthetic data is emerging as a scalable alternative, supplementing real-world datasets for improved model training.

4.1.2 LLM Generated Datasets

Based on extensive research, Wang et al. [93] proposed the LLM-Fake Theory from a social psychology perspective. This theory posits that news content generated by Large Language Models can be systematically categorized into two broad types: A. Fake News: a) style manipulation/sheeps clothing, b) content distortion, c) information fusion, d) narrative generation. B. Legitimate News: a) writing enhancement, b) news summarization. Building on this theoretical framework, they further developed the MegaFake dataset [93].

Notably, Lucas et al. [94] extended the generation strategies by innovatively incorporating perturbation-based generation and paraphrase-based generation, effectively simulating the heterogeneity of fake news. The F3 dataset was rigorously ensured the quality of the samples in terms of logical coherence, factual accuracy, semantic consistency, and contextual relevance. Huang et al. [95] further advanced these techniques by evaluating the effectiveness of ChatGPT through four distinct prompt engineering strategies. Their self-assessment and human evaluation results indicate that the quality of model-generated text is now comparable to human-written news. These findings confirm that optimized prompt design allows LLMs to produce high-quality samples for both real and fake news categories.

Similarly, the PROPANEWS dataset [96], constructed based on pre-trained models and containing 2,256 annotated samples, has shown significant practical value. Detection models

trained on this dataset achieved an F1-score improvement of 3.6% to 27.6% on public test sets, exhibiting particularly strong performance in manually written fake news detection tasks. These empirical studies provide compelling evidence that LLMs possess distinct advantages in generating highly deceptive texts and enhancing propagandistic rhetoric.

Recent work has focused on creating specialized datasets to address specific domains and challenges. Hu et al.[97], released a bilingual reasoning dataset generated via GPT-3.5. This resource supports analysis in both Chinese and English and facilitates the study of how explanatory text influences model performance. The ChatGPT-FC dataset, developed by Li [98], covers 154 topics related to politics, economics, taxation, and other domains from 2007 to 2023, with a particular emphasis on content associated with elections, healthcare, and social media information. This dataset facilitates multi-perspective and cross-domain disinformation detection. The SciNews dataset [99], consisting of 2,400 scientific news articles (both human-written and LLM-generated), utilizes jailbreak techniques to bypass model safety restrictions. The MR2-LLM multimodal dataset, developed by Wang [100], integrates raw data and LLM-generated explanatory texts.

The continuous emergence of these innovative datasets signals that LLMs are reshaping the research methodologies in the field of disinformation detection.

4.2 Data Augmentation

The detection of disinformation presents significant challenges due to the unstructured and unlabeled nature of data collected from social media platforms. Such data often contain large volumes of informally expressed raw content, making the direct extraction of structured information particularly difficult. Additionally, factors such as missing contextual information, low-resource languages, and class imbalance within training samples further hinder the effectiveness of disinformation detection models. As the dataset size increases, model performance tends to plateau, and the growing presence of noise and stylistic variability may

lead to a decline in precision.

A key challenge in data fusion lies in maintaining the overall semantic integrity of the dataset while independently transforming one modality or enhancing each modality without distorting the underlying information structure. This issue is particularly pronounced in multimodal disinformation detection, where balancing cross-modal consistency with content diversity is critical.

To address these challenges, data augmentation techniques are essential for improving both the efficiency and accuracy of disinformation detection models. These techniques facilitate the preprocessing of input data without losing critical information, thereby ensuring that the quality and usability of the information remain intact. Within the domain of disinformation detection, data augmentation can be categorized into five key components:

4.2.1 Text Normalization

Text normalization functions as a preprocessing mechanism designed to transform raw input into a standardized format. This process reduces the complexity of user-generated content and ensures that downstream classification models receive consistent data. Recent advancements indicate a transition from simple mechanical corrections to semantic refinement using generative models.

Social media platforms frequently contain slang and emotional language that obscure factual assertions. To address this, the CICAN framework employs ChatGPT to refine input data [101]. Through task-specific prompts, this method directs the model to correct linguistic errors and remove special characters while preserving original meaning. This approach effectively bridges the gap between noisy data and structured feature extraction.

Specific fact-checking tasks require more rigorous standardization known as Claim Normalization. The Check-worthiness Aware framework distinguishes itself from general summarization by employing Chain-of-Thought reasoning to extract verifiable statements [92].

Unlike general text summarization which captures the gist of a document, this technique reduces the cognitive load on human checkers by stripping away irrelevant details, although establishing universal standards remains difficult due to subjective editorial norms.

Normalization strategies also intersect with data augmentation to enhance model robustness in low-resource scenarios. For languages with limited datasets, such as Romanian, researchers utilize Back Translation and Easy Data Augmentation to generate synthetic samples [102]. These methods normalize data distribution by creating paraphrased variations which helps classifiers generalize without extensive manual labeling. Similarly, these techniques improve word embeddings by expanding the training corpus [103]. However, challenges persist including the high computational cost of translation services and the dependency on prompt quality where poor design can lead to inconsistent normalization outputs.

4.2.2 Textual Enhancement

Textual enhancement within data augmentation focuses on modifying the linguistic properties of news articles to create diverse and robust training samples.

A primary application of this technique addresses the scarcity of resources for specific languages or dialects. In the context of the Algerian dialect, researchers investigated the potential of translation-based augmentation to bridge the gap between high-resource languages like Modern Standard Arabic and low-resource dialects [104]. By employing Large Language Models such as GPT-4 to convert standard text into dialectal versions, this process demonstrated that automated data could substitute for manually curated datasets. Although this method improved recall by exposing classifiers to broader features, it introduced noise that occasionally reduced precision.

Beyond addressing scarcity, enhancement techniques mitigate the fragility of detection models against adversarial writing styles. Malicious actors increasingly utilize generative models to mimic the objective tone of reputable news, allowing fake content to bypass style-

based detectors. To counter this, Wu et al. [105] introduced a framework that reframes articles into various styles, such as converting sensational pieces into neutral ones. This approach forces the model to prioritize content consistency over linguistic tone. Furthermore, detection systems often fail to perceive fine-grained semantic changes. To address this, researchers proposed a framework utilizing semantic-flipped and semantic-invariant augmentation [106]. This method generates synthetic samples where meaning is negated or paraphrased. Integrating these samples into a contrastive learning objective ensures the system processes logical relationships rather than relying on textual patterns. Table 4.2 summarizes these augmentation strategies.

Task	**Task Introduction**	**Advantages**
Translation-based data augmentation [104]	Using automatic translation to generate synthetic training data for low-resource languages.	Addressed the scarcity of specialized datasets in regional dialects and reduced the high cost of manual data collection.
Semantic-based data augmentation [106]	Creating synthetic claims through text manipulation to help models understand underlying meaning.	Improved model resistance to adversarial attacks where slight linguistic changes like negations often mislead detectors.
Style-independent news reframing [105]	Using large language models to rewrite news in different styles while maintaining the same core information.	Mitigated detector vulnerability to stylistic mimicry where fake news adopts the objective tone of reliable publishers.

Table 4.2: Table of Textual Enhancement of FND

4.2.3 Cross-Modal Augmentation

Specifically, cross-modal augmentation primarily encompasses traditional techniques such as Optical Character Recognition (OCR) parsing, image captioning, and cross-modal alignment to achieve the association and unification of data across different modalities [107]. Cross-modal augmentation techniques address the semantic gap between visual and textual data by

enriching feature representations.

A primary strategy involves transforming visual content into textual formats to assist language-centric models. Wang et al. [100] introduced a pipeline that converts visual information into text using optical character recognition and image captioning. This conversion allows Large Language Models to interpret multimodal misinformation and retrieve external evidence. This method addresses the limited reasoning capabilities of standard models on raw visual data. Building on this concept of semantic conversion, Wu et al. [108] employed an image semantic enhancement module that generates captions from news images. Their approach integrates these captions with an efficient cross-modal prompt mechanism. By injecting complementary information during the early feature extraction stages, this method improves model adaptability across diverse news domains and reduces the reliance on heavy fusion networks.

While semantic translation bridges the modality gap, other approaches focus on expanding the training data through feature manipulation to address data scarcity. This technique is relevant for detecting fake news on emerging topics where annotated samples are rare. Jiang et al. [109] proposed a method for multi-modal fake news classification that requires only a small number of training samples. It achieved best results on three benchmark datasets, with an average accuracy improvement of 3.3% over the best-performing baseline.

Beyond increasing sample quantity, cross-modal augmentation also serves to refine the alignment between mismatched modalities often found in fake news. Standard contrastive learning often struggles with the ambiguity of image-text pairs in misinformation. To resolve this, Wang et al. [110] constructed an auxiliary dataset by generating mismatched image-text pairs from the original data. This strategy introduces a consistency learning task that uses these artificial negative pairs to create soft targets for the loss function, and improves the detection of false connections where images and text do not semantically match.

4.2.4 Data Expansion

Real-world datasets for fake news detection frequently exhibit a skewed distribution with significantly fewer fake news samples than real news articles [111]. This class imbalance impairs the ability of learning models to predict minority class examples accurately. Furthermore, existing repositories often lack sufficient social context information, such as user interactions and propagation patterns. To address these limitations, recent research utilizes large language models to synthesize diverse samples and simulate user reactions.

RumorLLM [112] addresses the data imbalance issue by capturing specific writing styles and content characteristics of rumors. By generating contextually relevant fake news samples, this method balances the dataset and improves detection accuracy without relying solely on limited real-world examples. While generating news content addresses textual imbalance, constructing a complete information ecosystem requires simulating user interactions. The DELL framework [113] introduces a diverse reaction generation component that simulates how different demographic groups perceive news articles. By defining user attributes such as gender, age, and political orientation, this method constructs a synthetic interaction network that reflects real-world complexities and biases. Extending the concept of user simulation, GenFEND [114] focuses on the challenge of silent users and early detection scenarios where real comments are unavailable. This framework uses large language models as user simulators to generate feedback based on specific profiles, aggregating these synthetic insights with available real data to enhance system performance during the initial phases of news propagation.

Data expansion techniques have also evolved to address the consistency required in multimodal detection. Jia et al. [115] utilize text-to-image models like DALL-E2 to generate and inpaint image regions based on textual prompts, ensuring strong semantic alignment between textual and visual modalities. Focusing on the quality rather than just the quantity of synthesized multimodal data, Ye et al. [116] employ semantic and distributional similarity

metrics to filter generated content. This selection process allows smaller multimodal models to achieve high performance in fact-checking tasks by training on high-quality synthetic data. We summarize these data expansion methods in Table 4.3.

Task	**Task Introduction**	**Advantages**
Diversified Sample Generation [112]	Minority Class Data Augmentation: Increasing data for smaller categories to balance the dataset.	The issue of class imbalance found in real-world fake news datasets.
Diversified Reaction Generation [113]	Simulating Interaction Networks: Large Language Models generate news reactions to represent diverse viewpoints and simulate user-news interaction networks.	Difficulty in obtaining actual user comments and reactions from social media platforms.
Comment Generation[114]	User simulators and comment generators.	It is difficult to obtain diverse comments in reality due to bias and different wills.
Multimodal Data Synthesis [115]	Automated Synthesis: Text prompt-guided image generation, combined with local masking and content reconstruction, to achieve automated data synthesis.	The problem of insufficient generalization ability in existing Large Language-Image models when generating locally tampered images.

Table 4.3: Table of Data Expansion of FND

4.2.5 Data Annotation

Obtaining high-quality annotated datasets remains challenging due to the labor required for manual verification. Consequently, researchers increasingly adopt automated and semi-automated techniques to reduce human effort while maintaining accuracy.

One approach leverages existing fact-checking resources to automate labeling. Akhtar et al. [117] proposed a method to construct large-scale ground truth datasets by matching social media posts with verified statements from sources such as Snopes and PolitiFact.

By using a BERT-based model to calculate semantic similarity, they successfully assigned binary labels (real or fake) to millions of tweets without human intervention. This approach addresses the scalability issue, allowing for the analysis of bot behaviors during crises like the COVID-19 pandemic. However, binary classification often fails to capture the complexity of disinformation which frequently blends truth with fabrication.

To overcome the limitations of simple binary labels, a multi-level annotation model tried to capture the semantic and structural dynamics of fake news [118]. Instead of a single truth value, their model categorizes content into seven distinct dimensions, including author intent, evidence presence, and disinformation technique. This granular approach treats detection as a structured information extraction task rather than simple classification. While this provides richer data, applying such detailed schemas manually is resource-intensive. Recent advancements in Large Language Models offer a solution to these constraints. Wang et al. [100] demonstrated that LLMs can serve as "teacher" models to automate the generation of detailed annotations. In their MMIDR framework, they utilized a teacher model to produce high-quality rationales and instruction-following labels for multimodal content. These machine-generated annotations were then used to distill knowledge into smaller, open-source student models. This strategy effectively bypasses the need for expensive proprietary models during deployment while ensuring that the detection system can provide interpretable explanations for its decisions.

Despite the capabilities of LLMs, their potential for "hallucination" necessitates mechanisms for verification. Li et al. [119] addressed this by proposing a hybrid "Self-Checker" framework that integrates automated generation with human oversight. In this semi-automated workflow, the system extracts claims and retrieves evidence to suggest initial verdicts, which are subsequently validated by human annotators. This human-in-the-loop approach combines the efficiency of machine generation with the reliability of human judgment, ensuring that complex texts are verified accurately against real-time evidence.

4.3 Information Extraction

Key information extraction plays a critical role in building efficient and reliable fake news detection systems. By extracting key features, core entities, events, and relationships from news texts, this process helps LLMs identify the primary and essential content of the information under scrutiny. This, in turn, enhances the efficiency of disinformation detection while also improving model interpretability by providing structured insights into the extracted content. This section introduces three main information extraction techniques: feature extraction, entity extraction, and claim extraction.

4.3.1 Feature Extraction

Feature extraction raw data into informative representations suitable for classification models. This phase determines the quality of the input signals, directly influencing the ability of downstream algorithms to distinguish between authentic and fabricated content. Capuano et al. [120] classify content-based features into linguistic, syntactic, style-based, and visual groups. Their review indicates that while deep learning models yield high accuracy, traditional methods like TF-IDF remain resistant to overfitting. Expanding on linguistic analysis, Madani et al. [121] propose an algorithm-level strategy to mitigate data quality issues, such as short text lengths and missing metadata. By measuring coherence and cohesion, their strategy identifies the disjointed narrative structures often present in manipulated content, effectively mitigating issues related to short text lengths.

To capture fine-grained logical fallacies, Wang et al. [122] apply Rhetorical Structure Theory to segment text into Elementary Discourse Units. This method models functional relationships, such as cause and elaboration, to reveal discourse patterns that differ between real and fake news. However, internal analysis struggles with semantic sparsity in short social media posts. Qiu et al. [123] introduce a Dual-layer Semantic Information Extraction Network (DSEN-EK) that incorporates external knowledge. Instead of relying on the provided text,

this method extracts entity descriptions from external sources like Wikipedia. By fusing context-based representations with knowledge-based features through a comparison function, the system ensures that the extracted features reflect verifiable facts, preventing the model from struggling with rare terms or limited context.

In multimodal environments, feature extraction must bridge the semantic gap between visual and textual data. Qi et al. [124] propose an entity-enhanced framework that extracts visual entities—such as public figures or landmarks—alongside textual entities. By treating visual content as high-level semantic features rather than raw pixel data, the system can calculate similarity scores and model the interaction between the two modalities. This allows for the detection of inconsistencies, such as a mismatch between the person mentioned in the text and the individual depicted in the associated image.

As manipulation techniques evolve, standard similarity scores often fail to capture subtle semantic conflicts, particularly in "cheap-fake" where images are paired with misleading captions. Wu et al. [125] advance feature extraction by utilizing LLMs not just for classification, but as feature extractors via prompt engineering. By querying an LLM to analyze specific relationships—such as subject consistency and contextual alignment—they generate high-dimensional feature vectors that represent complex reasoning. This method enables the detection system to identify contradictory relationships that standard embedding models might overlook. Table 4.4 summarizes representative methods in feature extraction, highlighting their specific objectives and advantages.

4.3.2 Entity Extraction

Entity extraction transforms unstructured news content into structured data by identifying key elements such as names, organizations, and locations. Early methods often treated all terms within a document with equal weight, but this leads to poor performance when dealing with short or domain-specific texts where context is sparse. To address this limitation, recent

Methods	Task Objectives	Advantages
EDUs [122]	Capture internal narrative logic and dependency structures	Facilitates early detection without reliance on external social context
EM-FEND [124]	Identify semantic inconsistencies between text and visual content	Bridges the semantic gap between general object labels and specific named entities
DSEN-EK [123]	Enrich short text representations with external facts	Overcomes semantic sparsity and ensures domain-specific understanding
Linguistic Feature Extraction [121]	Quantify the structural quality and logical flow of text	Provides robustness when processing datasets with limited metadata or quality issues
Cheap-Fake Detection [125]	Distinguish subtle relationships like contradiction versus entailment	Outperforms standard similarity metrics by leveraging advanced semantic reasoning

Table 4.4: Types of Feature Extraction of FND

research focuses on enriching entity representations through external knowledge bases. For instance, the Dual-layer Semantic Information Extraction Network (DSEN-EK) improves semantic understanding by retrieving descriptions of entities from sources like Wikipedia [123]. By comparing the news content with these external facts, the model assigns appropriate importance to different entities and verify consistency, ensuring the system processes specific vocabulary effectively rather than relying solely on the provided text.

While external knowledge solves issues related to semantic sparsity, social media environments introduce challenges regarding data quality and informal language. The irregularity of user-generated content, which often includes abbreviations and misspellings, hinders traditional feature extraction. To mitigate this, the CICAN utilizes large language models to refine raw text before processing [101]. This method extracts clean entities from noisy posts and constructs heterogeneous graphs to capture long-distance dependencies that standard

encoders miss. This progression highlights a shift toward using generative models as pre-processors that clean data and mine structural information from informal text. Similarly, research using PEFT/LoRA-based Fine-tuned Model demonstrates the ability to extract entities and their corresponding sentiments into structured formats like JSON [126]. This method uses specific instructions to guide the model, allowing for deep textual analysis and the identification of propaganda narratives on consumer-grade hardware. This represents a move toward parameter-efficient techniques that provide detailed entity analytics without the high resource costs associated with proprietary services.

Effective fake news detection often requires analyzing more than just text, particularly when misinformation relies on manipulating visual context. Multimodal strategies expand the scope of entity extraction to include visual data, addressing the limitation where text-only models fail to spot cross-modal inconsistencies. The Entity-enhanced Multimodal Framework addresses this by extracting visual entities, such as celebrities and landmarks, alongside textual entities [124]. This method bridges the semantic gap where generic object detection labels are too vague. By comparing the specific identities found in images with those mentioned in the text, the system calculates similarity scores to detect mismatches, ensuring that visual evidence supports written claims. Table 4.5 summarizes representative methods in entity extraction, highlighting their specific objectives and advantages.

4.3.3 Claim Extraction

Besides, claim extraction is also an important step in the information extraction process by transforming unstructured content into discrete units for downstream verification. While earlier research prioritized short or isolated claims, the generation of text by large language models requires methods capable of handling cohesive and lengthy paragraphs where factual assertions exist within complex sentence structures. Addressing this shift, recent frameworks employ the reasoning abilities of language models to decompose these long-form texts into

Methods	Task Objectives	Advantages
CICAN [101]	Extracting clean entities and abstract concepts from noisy social media posts to construct entity-sentence heterogeneous graphs.	Mitigates the noise inherent in informal text and captures complex long-distance dependencies using generative auxiliary tools.
PEFT/LoRA-based Fine-tuned Model [126]	Identifying named entities and associated sentiments to output structured data (JSON) for manipulation analysis.	Enables efficient, privacy-preserving textual analysis on consumer-grade hardware without relying on closed-source models.
Entity-enhanced Multimodal Framework [124]	Extracting specific visual entities (e.g., celebrities) and textual entities to calculate similarity scores and detect inconsistencies.	Bridges the semantic gap between images and text, allowing for the detection of specific entity mismatches in multimodal news.
DSEN-EK [123]	Retrieving and integrating entity descriptions from external knowledge bases (e.g., Wikipedia) to enrich short text representations.	Addresses semantic sparsity in short texts and ensures consistency by validating content against established knowledge sources.

Table 4.5: Types of Entity Extraction of FND

atomic and verifiable statements without extensive training [119]. This modular approach utilizes prompting strategies to isolate check-worthy segments from the surrounding context and prepares data for evidence retrieval. Although this method reduces the resources required for fine-tuning specialized models, it introduces dependencies on instruction design and increases latency due to the sequential processing of extracted units.

4.4 Model Validation and Refinement

LLMs have demonstrated the ability to assess the authenticity of information even without explicit reliance on external knowledge sources. By conducting in-depth contextual analysis, recognizing linguistic styles, identifying semantic associations, and extracting entities and

concepts, LLMs can capture implicit information and latent meanings to evaluate the credibility of a given text. Meanwhile, integrating external information and specialized tools has been shown to significantly enhance the accuracy of disinformation detection [26, 127, 119].

Despite their promising performance, LLMs still face several challenges in disinformation detection, particularly in areas such as handling complex semantics, validating evidence from multiple perspectives, and improving interpretability. To address these challenges, researchers frequently implement adaptive refinements, such as retrieval-augmented generation (RAG) techniques, fine-tuning on domain-specific datasets, incorporating adversarial training, and optimizing reasoning capabilities. Ultimately, once detection results are aggregated, a final decision-making process synthesizes all available judgments and extracted information, ensuring a comprehensive and well-informed assessment of the content's veracity.

4.4.1 Internal & External Validation

Traditional content-based fake news detection frameworks analyze textual content and writing style across lexical, grammatical, semantic, and discourse levels. In contrast, LLMs leverage internally acquired knowledge during pretraining, functioning as an internal search engine for verification [78]. Researchers have utilized models like ChatGPT to capture sequential and hierarchical features, integrating semantic structures for improved misinformation detection. This internal validation extends to multimodal analysis. Sniffer [26] identifies inconsistencies between images and text to enhance cross-modal analysis. It detects out-of-context misinformation by internally checking if the visual content logically aligns with the caption. Kim et al. [78] further refine internal processes through a multi-agent debate framework. This system uses agents as debaters and judges to review generated explanations iteratively. They ensure the reasoning remains faithful to the input evidence and reduce the risk of hallucinated details. This ability to verify information without external resources is valuable in emerging

domains where outside data is limited.

Beyond content-based approaches, external knowledge-based frameworks incorporate search engines, static and dynamic databases, and auxiliary tasks to enhance verification. LLMs serve a dual role, both as an intelligent search engine retrieving real-time information and as a language processor optimizing and filtering search results for improved accuracy. For instance, Self-Checker [119] exemplifies this by generating search queries based on specific statements. It forwards these queries to external engines like Bing to retrieve relevant evidence and verify the claims. Quelle et al. [128] utilize the ReAct framework to interact with APIs and fetch additional contextual data. Their research indicates that models utilizing external search capabilities consistently outperform those relying solely on internal training data. This external connection helps manage the rapid evolution of news and reduces errors associated with static parametric memory.

Advanced frameworks integrate these internal and external mechanisms into structured workflows. FACTOOL [129] employs auxiliary reasoning tasks to improve accuracy across multiple domains. It validates test cases for code generation and verifies step-by-step calculations for math problems by connecting with tools like Python interpreters and Google Scholar. FactAgent [86] simulates expert workflows by decomposing fact-checking into structured sub-tasks. It integrates internal knowledge for linguistic analysis with external tools for source verification to complete each step. This structured approach mimics human expert decision-making and reinforces the reliability of the system. Table 4.6 summarizes representative methods in Internal and External Model Validation, highlighting their specific objectives and advantages.

4.4.2 Model Refinement

Advancements in fake news detection using LLMs primarily focus on retrieval-augmented generation [130], RAG offers an effective paradigm for enhancing LLMs by addressing

Methods	Task Objectives	Advantages
Self-Checker [119]	External validation of LLM outputs using Bing search to verify claims.	Eliminates the need for fine-tuning; decomposes complex texts into checkable claims.
FacTool [129]	Detect errors in diverse tasks (text, code, math) using external tools like Python and Google Scholar.	Domain-agnostic framework; verifies high-stakes content beyond standard news text.
FactAgent [86]	Hybrid validation using internal consistency checks and external search for news verification.	Emulates human expert workflows; reduces reliance on large annotated training datasets.
Multi-Agent Debate Refinement [78]	Internal validation to ensure generated explanations align logically with provided evidence.	Improves trust by reducing hallucinations through iterative feedback and consensus.
Sniffer [26]	Validate image-text consistency and external veracity to detect out-of-context misinformation.	Provides natural language explanations; specifically tuned for recognizing news entities.
ReAct Framework [128]	Comparative validation of internal parametric knowledge versus external information retrieval.	Demonstrates the necessity of external context for reducing errors; assesses multilingual capabilities.

Table 4.6: Types of Internal and External Model Validation of FND

their inherent limitations in knowledge-intensive tasks [131, 132]. By integrating external knowledge retrieval, RAG mitigates issues such as factual hallucinations and the temporal lag in model parameters, as shown in Figure 4.2. This approach mitigates LLMs' limitations in knowledge-intensive tasks, ensuring more reliable fact-checking.

Retrieval Augmented Generation provides a direct source of external knowledge for verifying facts, moving beyond models that act merely as search engine proxies [133]. By retrieving evidence from databases, these systems effectively verify claims and generate labels, often outperforming larger parameters models [134]. Singal et al. [135] construct a three-stage system that connects dense document retrieval with few-shot in-context learning.

This approach extracts specific text segments to support veracity predictions. Building on the need for transparency, Niu et al. [136] introduce VeraCT Scan framework, which prioritizes justifiable reasoning by cross-referencing claims with internet-wide searches. This method addresses conflicting information by weighting evidence based on source credibility rather than treating all retrieved text equally.

While standard retrieval provides a foundation for verification, static strategies often fail when ensuring coverage of complex or rapidly evolving online content. Li et al. [137] proposed the STEEL framework, which utilizes an adaptive mechanism that evaluates evidence sufficiency and initiates additional search rounds if confidence remains low. Increasing the complexity of reasoning further, Khaliq et al. [138] apply iterative retrieval to multimodal claims through their RAGAR framework. This method decomposes political claims into sequential questions using a tree-based reasoning structure. It analyzes image captions alongside text to synthesize a final verdict from multiple retrieval steps. Morever, Bai et al. [139] tackle hallucinations where no evidence exists with the Full-Context Retrieval and Verification framework. Their method treats the absence of retrieved evidence as a distinct signal of falsification rather than a failure of the search engine. This allows the system to identify machine-generated fabrications that lack a basis in reality. Beyond detection, RAG techniques assist in hardening models against sophisticated attacks. Singh et al. [140] utilize RAG to generate contextually grounded fake news for adversarial training. By simulating high-quality misinformation, they refine the robustness of the detection pipeline against

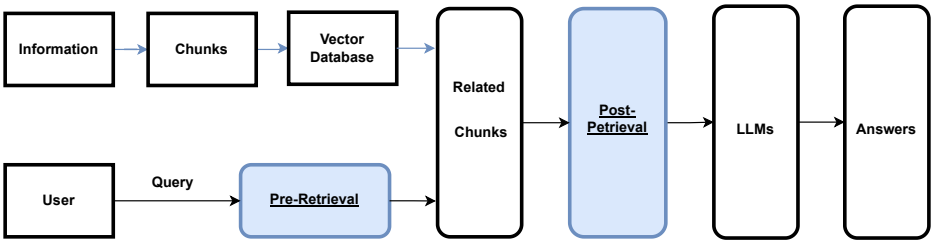

Figure 4.2: Basic Pipeline of RAG

linguistically fluent but factually incorrect text.

Despite the benefits of external context, RAG systems can underperform zero-shot baselines when exposed to noisy results [141]. To mitigate reliance on potentially unstable training data, Nezafat et al. [142] propose a training-free approach using open-source models like Mixtral combined with chain-of-thought prompting. High-stakes domains require even stricter adherence to accuracy. Li et al. [143] develop agentic architectures, such as Corrective RAG, which actively critique retrieval quality and rewrite queries if initial evidence proves insufficient. Complementing this, Upadhyay et al. [144] refined extraction by prioritizing factual consistency over topical relevance, generating trusted baselines from scientific databases to prevent the propagation of misinformation during the retrieval phase.

Other methods use reranking to reduce the semantic gap between user queries and documents [145]. These techniques allow the model to choose the best evidence and improve detection results. RAG also demonstrates effective performance in cross domain and multimodal detection through specific reasoning strategies [146]. Certain models use a chain or tree of questions to retrieve evidence in a step by step manner. This approach is useful for complex fields such as politics [147].

4.4.3 Final Decision

Researchers typically combine pre-trained knowledge from large language models (internal judgment) with external knowledge and tools (external judgment) in the final decision stage of the LLM-based FND framework to improve the reliability and accuracy of the overall judgment [86, 99]. For example, Liu [148] propose a decision module that merges results from internal (pretrained LLM-based) and external (knowledge/tool-assisted) judgments. If both align, the model directly adopts the result with high confidence. Otherwise, it conducts further analysis, integrating predictions and explanations before making a final decision. Similarly, certainty modules evaluate aligned or conflicting results, refining assessments

through a structured verification process.

4.5 Results Explanation and Feedback

LLMs have introduced new paradigms in fake news detection, but their complexity and "black-box" nature have raised concerns regarding explainability [149, 150]. Improving model interpretability not only enhances user trust but also helps researchers identify biases and anomalies in data, facilitating model refinement and performance enhancement.

One of the key advantages of LLMs over traditional methods is their ability to provide justifications for classification decisions. LLMs can generate interpretable reasoning rules, allowing researchers to adjust these rules manually for better adaptability and control. This transparency ensures that detection frameworks remain relevant and accurate in combating misinformation. Studies on ChatGPT's interpretability have demonstrated its capability to explain fake news classifications [95], while argumentative LLMs use structured reasoning frameworks to enhance decision transparency. Systems like OE-Fact [127], FOLK [151], and FactAgent [86] provide real-time justifications for fact-checking, reducing opacity in decision-making. Moreover, frameworks such as MMIDR [100], STEEL [137], SNIFFER [26], and EXMULF [24] extend explainability to multimodal misinformation, ensuring coherent analyses across text and images. Methods like HiSS [152] and MADR [78] enhance user trust by making each reasoning step traceable and verifiable.

Feedback mechanisms play a crucial role in improving LLM-based fake news detection by continuously refining model predictions. These mechanisms include human feedback alignment, agent-based feedback loops, and adversarial training. Fine-grained feedback, which provides sentence- or paragraph-level corrections, enhances LLM performance by improving training signal density and eliminating irrelevant or inaccurate information. Automated feedback, such as ULTRAFEEDBACK [153], addresses scalability issues by replacing human annotations with large-scale AI-generated critiques, reducing labeling costs while improving

model alignment. Besides, ExFake [30] is composed of four modules working together to return the percentage of confidence of an input tweet and provide interpretable explanations to OSN users, it outperforms all the state-of-the-art baselines of the FakeNewsNet benchmark on all the metrics in experiment.

A novel approach integrates external knowledge with feedback mechanisms to mitigate hallucination in model outputs. By leveraging external verification sources, this method ensures content accuracy while using automated feedback for refinement. Additionally, self-generated feedback reduces reliance on extensive supervised datasets. The Deception Detection framework employs a three-stage process—initial prediction, LLM-based review, and correction—to iteratively enhance prediction accuracy [154]. The MADR framework further improves explainability by utilizing multi-agent debate optimization, where multiple LLM agents refine explanations collaboratively [78]. This multi-agent strategy is expected to enhance transparency and credibility in future fake news detection systems, reinforcing public trust in AI-driven fact-checking solutions.

Chapter 5

Experimental Results and Analysis

Chapter 3 underscores the evolving landscape of fake news detection, as the field increasingly leans towards sophisticated computational methods. However, several critical gaps remain in the practical application of these technologies.

Firstly, computational complexity and resource intensity are major barriers. State-of-the-art DL models require substantial GPU resources for training and inference, making them difficult to deploy in resource-constrained environments or on edge devices (such as mobile phones). Secondly, the "Black Box" nature of Deep Learning poses a challenge regarding interpretability. In the context of fighting disinformation, understanding why a piece of news is flagged as fake is almost as important as the detection accuracy itself. Complex neural networks often lack the transparency required for trust by human moderators. Finally, previous studies suggest that on smaller, topic-specific datasets, the performance gain of complex DL models over robust classical Machine Learning (ML) models is often marginal. Therefore, it is critical to re-evaluate the efficacy of classical ML algorithms combined with effective feature engineering.

This chapter outlines the methodology employed in the case study on Machine Learning-Based Fake News Detection. Given the complexities and the multifaceted nature of fake news

dissemination, especially during significant events like the 2016 U.S. presidential election, our approach combines meticulous data collection, exploratory data analysis, and the application of various machine learning classifiers [155], [156]. The methodology is structured to not only identify linguistic patterns and features unique to fake news but also to compare the accuracy of fake news detection [157].

5.1 Dataset

The dataset used in this thesis is from the Kaggle website and features news reported by nine different news organizations in the week leading up to the 2016 U.S. election. The content of the data has been certified by Buzzfeed to be authentic. The dataset is divided into two categories: fake news and real news. Each category includes 91 entries and 12 different attributes. After checking, the variables are described as strings and there are no missing values in the dataset which indicates that the dataset is clean and well structured for further analysis. The attributes which include ID, title, text, source, images, and videos [158], [159].

- **id:** A unique identifier for the webpage of the news article, indicating its authenticity.
- **title:** The headline designed to capture reader's interest, is closely related to the news topic's essence.
- **text:** The article's main body, detailing the news story and often emphasizing and elaborating on a central claim.
- **source:** The author or publisher of the news piece.
- **images:** Visual elements that support the article's content, aiding in story framing.
- **videos:** Video content embedded in the news article, including video clips of the news story or related footage.

5.2 Text processing

Data preprocessing is a key step in the machine learning workflow, which directly affects the performance of the model, training efficiency and generalization ability. This part mainly relies on the `preprocess_corpus()` function. to clean the dataset and do the transformation and integration for different attributes, the specific steps are as follows:

1. **Lowercasing:** Converts all text to lowercase to ensure uniformity across the corpus.
2. **Removing Numbers:** Deletes any numeric characters, as they are typically not relevant to the analysis of news authenticity.
3. **Eliminating Punctuation and Special Characters:** Removes punctuation and specific special characters (e.g., '<', '...') that are irrelevant for text analysis.
4. **Excluding English Stopwords:** Removes common English stopwords to highlight more meaningful words within the text.
5. **Removing News Source Names:** Excludes common news source names from the corpus to prevent bias in the analysis.
6. **Applying Stemming:** Reduces words to their root forms, facilitating a more consistent analytical approach.
7. **Removing Extra Whitespaces:** Cleans up the corpus by eliminating superfluous whitespaces.

The implementation of these text cleaning steps is encapsulated within the `clean_text` and `preprocess_corpus` functions, as outlined in the provided code snippet. This meticulous preprocessing ensures the standardization of the text data, preparing it for detailed analytical procedures.

Following the cleaning process with the `preprocess_corpus()` function, the analysis advances to identifying words that are distinctly associated with either 'real' or 'fake' news categories. A chi-square test is conducted to determine the statistical significance of the occurrence of specific words within each category. This step aims to uncover patterns or indicators that could assist in distinguishing between real and fake news.

Besides, we implemented a comprehensive URL normalization process to minimize potential discrepancies that could arise from inconsistent URL formatting. This step was crucial in enhancing the reliability of our analysis, allowing for a more accurate evaluation of the credibility of the news source, free from the biases introduced by URL inconsistencies.

In addition, when there is an image or video link under the entry, the corresponding attribute will be assigned a value of 1, and conversely if there is no link under the attribute it will be assigned a value of 0. BEesides, the original format (URL) of images and videos has been innovatively transformed into categorical variables, making abstract content easier to handle. By integrating disparate datasets, introducing essential variables for differentiation, refining variable representations, and conducting advanced textual analysis, a comprehensive dataset is prepared. This dataset is primed for in-depth analysis in subsequent phases of the study, aimed at unraveling the nuances of news authenticity.

5.3 Exploratory Data Analysis

Our EDA focused on understanding the distribution of key features within the dataset, identifying patterns and characteristics unique to fake and real news articles. This involved analyzing term frequencies in news titles and bodies, examining title lengths, and exploring the presence of specific words and phrases (unigrams and bigrams) that might indicate the authenticity of a news article [160], [161].

5.4 Machine Learning Models for Classification

This section explored several machine learning models to develop a robust classifier capable of distinguishing between real and fake news effectively. The following models were employed:

5.4.1 Multi-Layer Perceptron (MLP)

The model is compiled using the 'RMSprop' [162] optimizer and the binary cross-entropy loss function. Accuracy is chosen as the evaluation metric. Training follows 5-fold cross-validation, reducing the learning rate on a plateau to optimize performance. Performance metrics, including loss, accuracy, precision, recall, and F1 score, are reported for each fold, with averages calculated across all folds.

5.4.2 Naive Bayes Classifier

Naive Bayes is a probabilistic classifier based on Bayes' theorem, assuming feature independence. It is trained using the naive Bayes function from the e1071 package in R, performing well in many text classification tasks [163]. The accuracy of the Naive Bayes model is assessed by comparing the predicted class labels with the true class labels on the test data, achieving an accuracy of approximately 54%.

5.4.3 Random Forest Classifier

The Random Forest classifier is an ensemble method that constructs multiple decision trees during training. By averaging the predictions of multiple decision trees, Random Forest reduces the risk of overfitting, improving generalization to unseen data [164]. In this implementation, 500 decision trees are trained, with a subset of features used at each split. The out-of-bag (OOB) error rate is approximately 26.47%, and final predictions are made through majority voting among the trees.

5.4.4 Logistic Regression Classifier

Logistic Regression is a linear model used for binary classification, making it well-suited for fake news detection, which involves determining the truth or falsity of news items. Additionally, it outputs probabilistic assessments, which help quantify uncertainty in predictions [165]. In this implementation, logistic regression is performed using the glmnet function with elastic-net regularization to prevent overfitting and improve generalization. Predictions are made using the logistic function, and class labels are assigned based on a threshold (usually 0.5). Model accuracy is evaluated by comparing predicted labels with actual labels on the test data.

5.5 Results and Interpretation

This section compared various machine learning models to distinguish between real and fake news articles, and assessed the performance of these models based on accuracy, precision, recall, and F1 score [73]. Below, we present our key findings:

5.5.1 Exploratory Data Analysis

Exploratory Data Analysis is a crucial component in research, as it provides reliable guidance for subsequent hypothesis validation and model construction. Previous studies have highlighted that the dissemination of fake news is significantly biased, with sources predominantly concentrated in non-mainstream and social media platforms. This dissemination pattern differs markedly from that of real news. Therefore, the primary objective of conducting EDA in our study is to identify the differences between fake and real news by examining their sources, content complexity, and distinctive characteristics.

We first examine the sources of all news items. As illustrated in Figure 5.1, both real and fake news sources are displayed. A closer analysis of fake news sources reveals that outlets

such as rightwingnews.com and eaglerising.com show a clear preference for fake news, with the number of fake news articles surpassing that of real news. In contrast, politi.co and cnn.it publish the most real news. It is also important to note that some sources of fake news were not labeled. Nevertheless, we retained this portion of the data as it offers additional

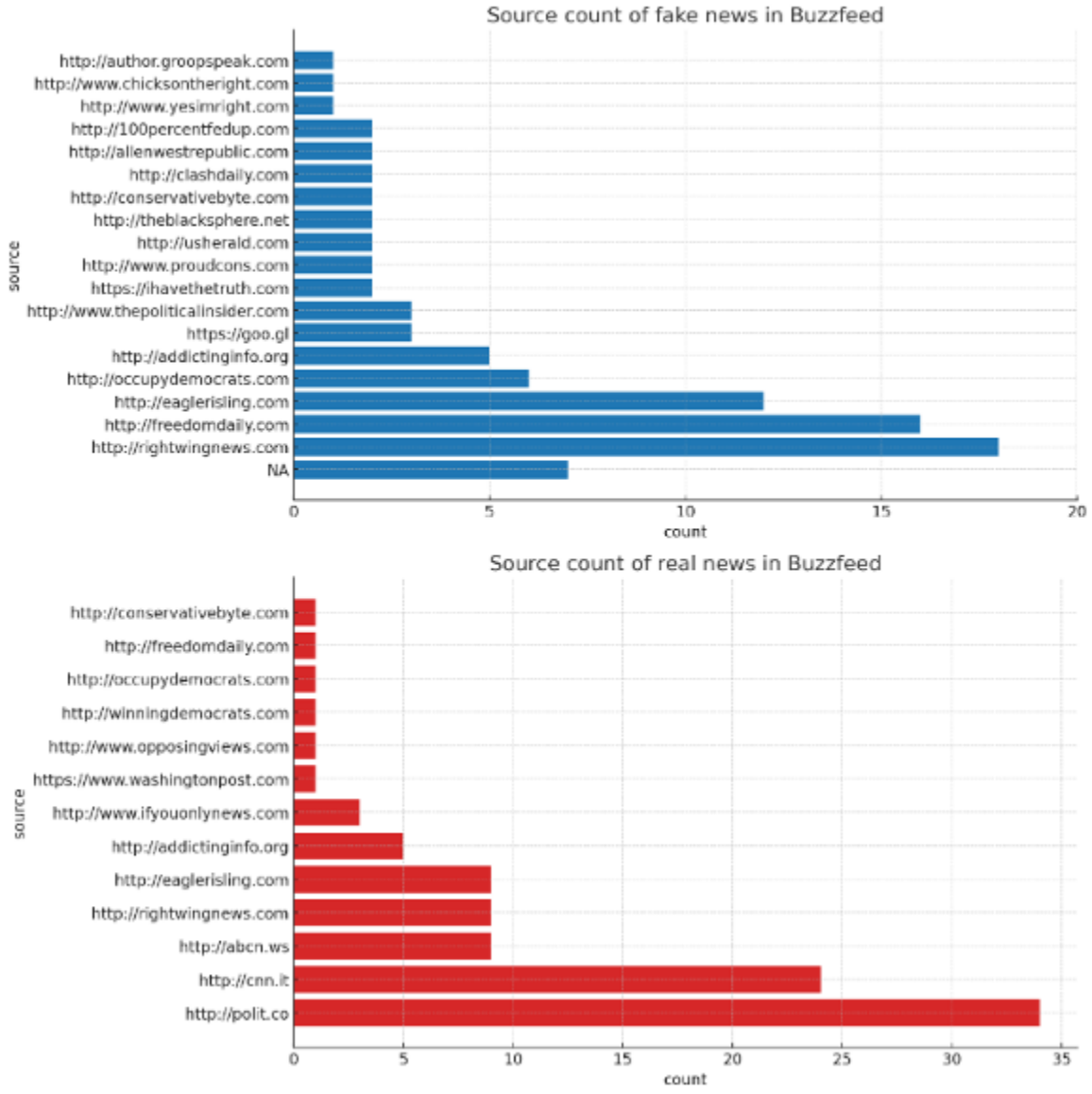

Figure 5.1: Statistical Bar Chart of Real and Fake News Sources

insights. While the sources of some fake news remain unidentified, all real news originates from reputable outlets familiar to the general public.

Additionally, it is essential to compare whether a given source produces both real and fake news. As shown in Figure 5.2, we identified eight common sources that publish both types of content, with fake news often being more prevalent. Sources like rightwingnews.com and eaglerising.com clearly favor fake news, whereas freedomdaily.com almost exclusively produces fake news with minimal real news reporting. Conversely, sources such as occupydemocrats.com and conservativebyte.com exhibit a more balanced or smaller distribution between real and fake news.

Figures 5.3 and Figure 5.4 show the distribution of the most discriminating words in news headlines and body text across real and fake news, providing important clues to key parts of the exploratory data analysis and informing further diagnosis of real and fake news using AI models.

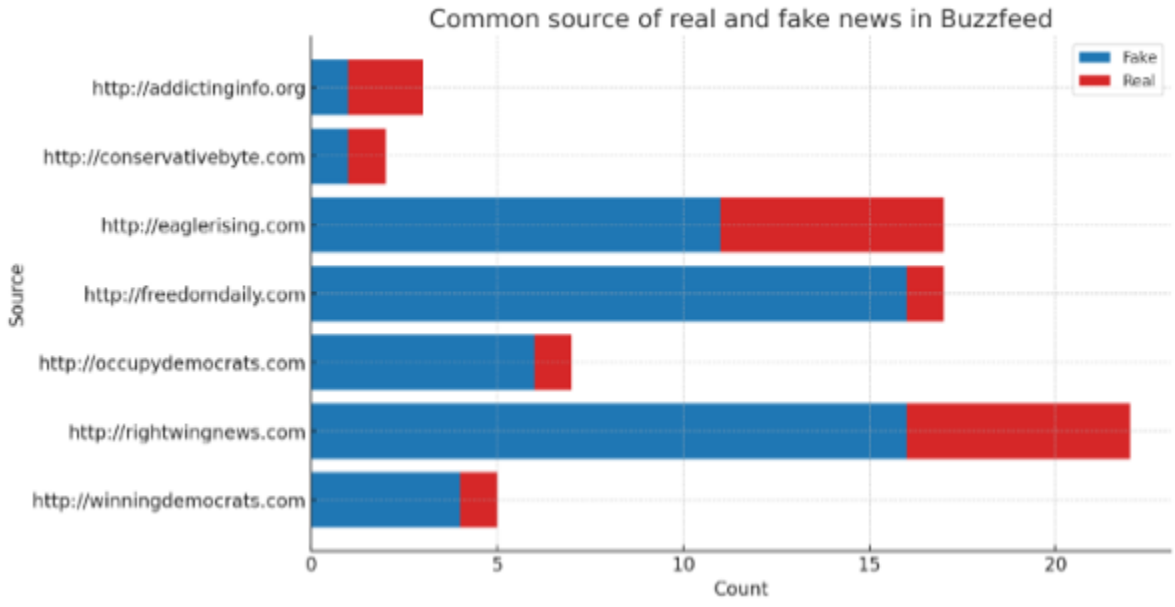

Figure 5.2: Comparison Chart of Common Real and Fake News Sources

Figure 5.3 illustrates the word frequency distribution of the 20 most discriminating words in news headlines. The blue bar represents the word frequency in fake news, while the red bars represent real news. The results indicate that the word “Trump” is prevalent in both fake and real news, but its frequency in fake news is significantly higher, appearing nearly 40 times. This suggests that Trump-related topics are frequently leveraged in fake news, likely due to their controversial nature, which tends to attract more attention. Other words such as “Hillary” and ”Muslim” also appear more often in fake news, highlighting a preference for political and religious topics. In general, fake news tends to use more inflammatory terms (e.g., “bomb,” “refuge”, “shoot”), while real news employs more neutral terms like “debate” and “first”. This difference suggests that fake news often relies on emotionally charged language to capture audience attention.

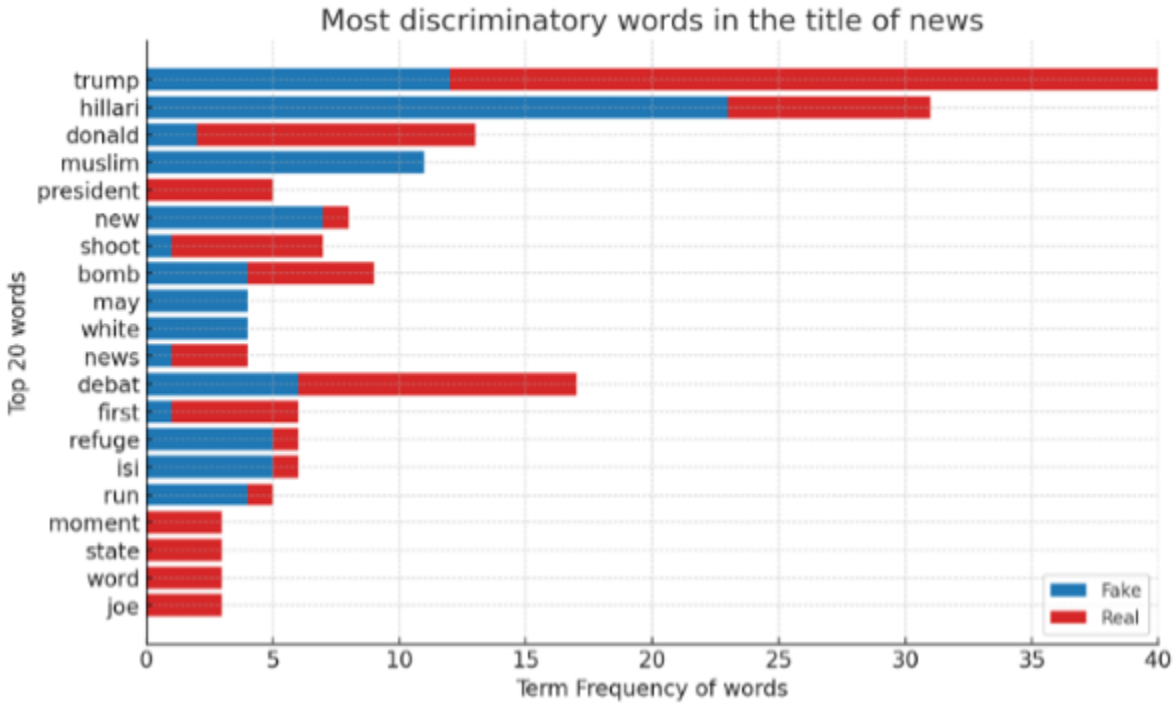

Figure 5.3: High Frequency Words in Fake and Real News Titles

Further analysis of the 30 most frequent words in the body of news articles shows a more concentrated distribution compared to headlines, especially when comparing fake and real news. As in the headlines, "Trump" and "Hillary" were also the most frequently mentioned keywords in both fake and real news. In the body text, "Trump" has a word frequency of up to 600 times, showing that Trump-related stories are the main carriers of false information. Moreover, the analysis reveals distinct differences in the emotional tone and stance words used in fake versus real news. For instance, neutral words like "said," "think," and "like" are more common in real news, reflecting its tendency to present facts and logical arguments. In contrast, fake news frequently uses emotionally manipulative terms such as "bomb" and "refugee." Additionally, temporal words like "Monday" and "Wednesday" appear more frequently in real news, likely because real news often references specific, verifiable timeframes, whereas fake news tends to rely on vague, unverifiable statements.

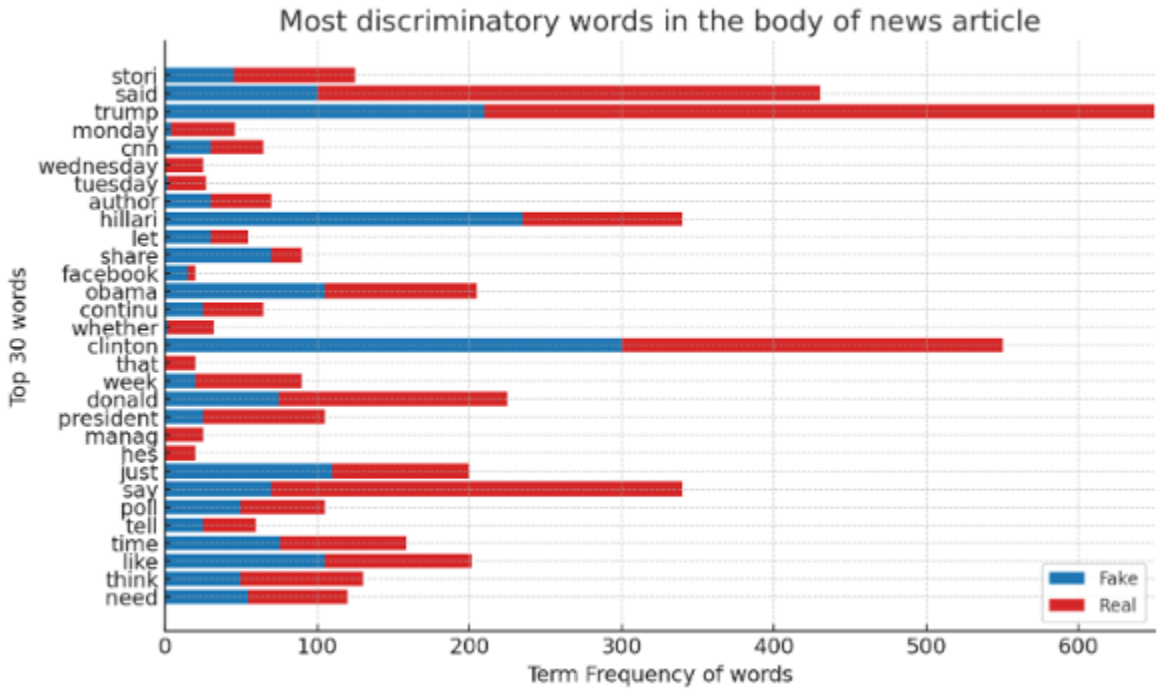

Figure 5.4: High Frequency Words in Fake and Real News Bodies

Finally, we conducted an in-depth analysis of the most frequent bigrams in the news text, comparing their occurrence in both real and fake news. This analysis highlights significant linguistic differences between the two types of content. Notably, "Donald Trump" is the most common double-word phrase in real news, while "Hillary Clinton" is a high-frequency phrase in fake news. This suggests that while news about these political figures is common across both types of stories, the contexts and narratives surrounding them differ. Some phrases, such as "Young Adults," "Clinton Foundation," and "Down Hawkins," are present in both real and fake news but are more frequent in fake news, indicating that fake news may exaggerate or distort certain topics to appeal to specific audiences or evoke emotional reactions. On the other hand, phrases like "Barack Obama," "New York," and "United States" are much more common in real news, suggesting that real news is more focused on actual events, people, and places. The lower frequency of terms like "Story Continued," "Barack Obama," and "Book 101" in fake news further suggests that real news is more likely to cover complex or in-depth topics, while fake news often focuses on simpler, attention-grabbing subjects.

5.5.2 Model Performance

This section evaluates four classical machine learning classifier models, including Simple Bayes, Random Forest, Logistic Regression, and Multi-Layer Perceptron (MLP), to identify fake news. These models are compared and evaluated using the main performance metrics (Accuracy, Precision, Recall, F1-Score) presented in Table 5.1 and the confusion matrix and accuracy scores of the four classifiers are extracted. By combining these evaluation metrics, it facilitates a comprehensive understanding of the performance of each model in handling the task of fake news detection.

The Multilayer Perceptron (MLP) performs the best overall among the four models, achieving an accuracy of 90.9%, a precision of 0.833, a recall of 0.556, and an F1 score of 0.667. These results indicate that MLP is able to effectively capture the subtle differences between

fake and real news, and its high precision rate and overall F1-Score make it show strong potential in text categorization tasks despite its relatively low recall rate. Therefore, MLP is able to strike a good balance between precision and recall, and shows strong robustness especially when dealing with the task of detecting false news.

Random Forest also shows good performance, with an accuracy of 83.6% and an F1 score of 0.608, indicating that it has a balanced performance in capturing positive and negative samples. The Random Forest model slightly outperforms the logistic regression in terms of the balance between precision and recall.

The logistic regression also performs well with 76.4% accuracy and an F1 score of 0.519, demonstrating its stability in the task of detecting false news. The model performs reasonably well in terms of recall (0.538), but is relatively weak in terms of precision (0.5).

The performance of the plain Bayesian model is weaker, with an F1 score of only 0.387 and an accuracy of 65.4%, indicating that the model has some limitations when dealing with fake news detection. Since the assumption of plain Bayes is premised on the independence between features, and the features in actual linguistic data tend to have strong dependencies, the model performs poorly in capturing the subtle differences in fake news. Moreover, the performance of plain Bayes is limited when confronted with complex text categorization problems, although it performs moderately well in certain simple categorization tasks.

Classifiers	**Accuracy**	**Precision**	**Recall**	**F1- Score**
Naive Bayes	0.654	0.429	0.353	0.387
Random Forest	0.836	0.583	0.636	0.608
Logistic Regression	0.764	0.5	0.538	0.519
Multi Layer Perceptron	0.909	0.833	0.556	0.667

Table 5.1: Performance Comparison of Various Classifiers

Chapter 6

Challenges, Future Trends and Conclusions

6.1 Bridging Theory and Practice: Efficiency vs. Complexity

The important investigation of this thesis involved contrasting the theoretical trends and key techniques observed (Chapter 3) with the empirical results obtained from the experimental implementation (Chapter 5).

Chapter 3 indicated a prevailing academic focus on increasing model complexity, with Transformer-based models often cited as the gold standard for accuracy. However, our experimental analysis challenges the notion that complexity is a prerequisite for effective detection. As demonstrated in Chapter 5.5.2, the implementation of The Multilayer Perceptron classifiers, achieved an accuracy of 90.9%. This performance is comparable to several complex baselines discussed in the review, yet requires a fraction of the computational time and power.

This finding bridges the gap between theory and practice by validating that in practical, real-world scenarios where computational resources are limited (e.g., real-time monitoring

streams), well-optimized classical Machine Learning models remain highly effective. The results suggest that the quality of feature engineering (e.g., text preprocessing and vectorization) often plays a more decisive role in classification performance than the sheer depth of the model architecture. Thus, while the theoretical trend moves toward Deep Learning, practical implementation warrants a continued reliance on and refinement of classical ML approaches for their explainability and efficiency.

From the experiment results, we conducted detailed exploratory data analysis on a Buzzfeed dataset containing both real and fake news. Our analysis included generating multiple plots for all variables within each news category. we meticulously explored both unigrams and bigrams to identify distinctive words and phrases commonly found in fake news articles, focusing on both titles and bodies of text. For unigrams, we employed a comprehensive text cleaning process. This included converting all text to lowercase, removing numbers, punctuation, selected special characters (like '<', '...'), and filtering out English stopwords and common news source names to diminish noise. Furthermore, we applied stemming to reduce words to their root forms, thereby simplifying our analysis to more generic language patterns.

For bigram analysis, we took a different approach, recognizing the importance of preserving specific word combinations that may convey more complex meanings, especially in the context of manipulative language often found in fake news. As a result, we opted not to apply the same rigorous cleaning process. We refrained from removing stopwords and stemming in order to retain the linguistic nuances that could be essential in identifying tactics employed in fake news. This decision allowed us to capture a more accurate representation of how language is manipulated in these articles, providing a deeper understanding of the subtleties behind fake news narratives.

In conclusion, the experiment and results showed that lighter, interpretable, and computationally efficient ML models can still achieve competitive performance benchmarks, offering a pragmatic alternative to resource-heavy deep learning frameworks.

6.2 Challenges and Future Trends

While existing works have established a solid foundation for AI-driven fake news detection research, further opportunities remain.

6.2.1 Data Evolution and Quality

The field of fake news detection has transitioned through distinct stages of data development and faces significant obstacles related to the quality of training information. Early research focused primarily on text datasets which represented fake news as a simple binary classification problem [166]. As communication platforms changed researchers began to include multiple types of media such as images and videos to reflect modern consumption patterns. This evolution from textual content to multimodal sets allows models to identify inconsistencies between different forms of media like a misleading headline paired with an unrelated image. Recent efforts have further expanded these resources to include video centric data and social interaction signals such as user comments and propagation patterns. These additions help capture how misinformation moves through networks but also introduce higher levels of complexity in data collection and management.

Despite these advancements several persistent challenges affect the performance of detection algorithms. Data imbalance remains a central issue because the volume of legitimate news typically far exceeds the amount of confirmed fake news available for study [167]. When models train on such imbalanced sets, they often develop a bias toward predicting the majority class which leads to a higher rate of false negatives. Noise within the data also complicates the learning process. Public datasets frequently contain typos or informal language from social media that can confuse standard processing techniques. Furthermore, the static nature of many established datasets presents a significant hurdle. Because the themes and methods of misinformation change rapidly a dataset created a few years ago might not contain the patterns needed to identify current deceptive narratives. This temporal decay means that

models may lose their accuracy over time as they encounter new forms of content they were never trained to recognize.

Bias in data collection and annotation further reduces the reliability and reproducibility of detection systems. High-quality datasets are essential for developing reliable machine learning models, while many existing datasets are outdated or imbalanced, which limits their effectiveness for training modern detection systems [168]. Selection bias occurs when researchers gather data only from specific topics like politics or certain types of websites while ignoring areas such as health and education [32]. This narrow focus limits the ability of a model to work well in different domains. Labeling bias is another concern because human annotators often have different interpretations of what constitutes fake news. Inconsistent labels can lead to contradictory training signals which prevent the model from learning clear distinctions between true and false information [168]. Many current resources are also limited by language as the majority of high-quality datasets are written in English. This lack of linguistic diversity makes it difficult to deploy effective detection systems in non English speaking regions or in multicultural environments [32].

To address these quality gaps researchers are exploring several practical strategies. Data augmentation serves as a key method to balance datasets by artificially creating more examples of fake news [169]. Techniques such as using generative models or Large Language Models can produce synthetic samples that improve the diversity of the training set without the high cost of manual collection. There is also a move toward creating dynamic datasets that receive regular updates from live social media feeds to keep the information current. Developing centralized portals to share and verify datasets helps the research community maintain consistent standards and reduces the duplication of effort. By focusing on these improvements researchers can build more robust systems that are capable of handling the evolving nature of misinformation across different languages and media formats.

6.2.2 Interpretability & Ethical Concerns

The deployment of automated systems for fake news detection introduces significant challenges regarding model interpretability and ethical responsibility. While deep learning models achieve high accuracy on benchmark datasets, their internal mechanisms often remain opaque. This lack of transparency creates a barrier to public trust and raises concerns about the potential for algorithmic bias. Current research indicates that when models operate as black boxes, they may rely on unintended correlations rather than factual verification. These issues necessitate a shift toward systems that are both effective and explainable to ensure they can be used safely in real-world scenarios.

Interpretability serves as a bridge between high-performance neural networks and the requirement for human oversight. In the context of journalism and law enforcement, a simple classification label is often insufficient. Users require justifications to understand why a specific article is flagged as false. In research area, reproducibility is fundamental to scientific validation [170], while interpretability is the precondition of reproducibility. However, many studies fail to provide clear algorithms or notations, and only a small fraction of reviewed papers offer publicly available datasets with thorough annotation guidelines. Challenges in this area include the trade-off between the complexity of architectures like BERT or Long Short-Term Memory networks and the clarity of their outputs [65]. Current trends involve the development of hybrid models that combine deep learning with probabilistic topic modeling [171, 172]. These approaches allow the system to extract semantic representations while providing a topic distribution that humans can understand. By integrating these methods, developers can offer evidence-based results that highlight specific words or themes influencing a decision, thereby reducing the intellectual isolation caused by opaque filtering algorithms.

The research examined ethical issues by implementing and testing three detection models to explore problems like algorithmic bias and generalizability. It highlights that the potential harm caused by automated tool: a specific finding is that ground truth labels can be heavily

skewed against certain political leanings which causes the models to systematically flag specific sources as unreliable regardless of the actual content [173]. Besides, ethical issues in fake news detection often stem from the data used to train these systems. Many models demonstrate a failure of the independent and identically distributed assumption, meaning they perform well on training data but struggle with news content from different time periods or topics. For instance, a model trained on political news may fail when applied to health-related misinformation, due to representation bias [173]. Therefore, there is an increasing push for multidisciplinary inclusion where experts from fields like sociology and psychology work with computer scientists to assess the ethical concerns that inevitably arise in AI model training. Moreover, the interpretable machine learning can mitigate fake news problems through transparent news feed algorithms and interpretable detection methods. By this way, it addressed ethical concerns such as unintentional discrimination, loss of opportunities, and social stigmatization that can result from biased algorithms in information distribution [174].

The research community is increasingly focusing on explainable AI as a standard component of detection tools [65]. One prominent trend is the use of model-agnostic surrogate techniques, such as Local Interpretable Model-Agnostic Explanations and Anchors. These methods function as plug-and-play modules that provide local explanations for individual sentences without requiring access to the internal parameters of the underlying model. Another developing direction is the creation of three-dimensional interpretability frameworks that address algorithmic transparency, human understandability, and the provision of supporting evidence from external knowledge graphs [174]. These advancements aim to move the field beyond a leaderboard culture focused solely on accuracy. Instead, the focus is shifting toward a multidisciplinary approach that incorporates sociology and psychology to evaluate the broader societal impact of automated news verification.

Ensuring the long-term reliability of these systems requires an ongoing process of model refinement and feedback. As social and cultural norms evolve, the definition of what constitutes fake news may change, making static datasets less effective. Future systems must incorporate

feedback loops where model performance is constantly re-evaluated against new information [173]. This process includes the use of soft moderation techniques, such as warning labels and media literacy reminders, rather than immediate content removal. By providing users with transparent news feeds and visualizations of feature weights, these systems can help eliminate filter bubbles and encourage more critical information consumption. The goal of these integrated strategies is to create a secure environment where AI supports human decision-making without compromising ethical standards or transparency.

6.2.3 Multimodal Complexity

The challenge of multimodal complexity in fake news detection arises from the increasing use of diverse media formats to spread misinformation. Most early research focuses on text-based analysis, yet modern fake news often combines images, videos, and social network data to create misleading narratives. This reliance on a single data type limits the ability of a system to identify inconsistencies between different formats. For example, a true text description might be paired with a manipulated image to change the overall meaning. Detecting such cases requires a system to analyze and compare multiple data sources at the same time. Integrating these different data types into a single framework presents a significant technical problem [175]. Each format requires specialized computational methods like natural language processing for text and computer vision for images. A successful model must go beyond simple feature combination and instead achieve a deep alignment between these modalities. This alignment is necessary to bridge the semantic gap where the same concept is represented differently in text and visual data. Current efforts are still in the early stages, and developing models that can synthesize information from these disparate sources remains a difficult task for the scientific community [176].

The lack of diverse and multilingual datasets further complicates the development of robust multimodal models [177]. Many existing datasets are limited to English and focus

on specific topics like politics. This bias makes detection systems less effective in other regions and cultures where misinformation might follow different linguistic patterns. Each language carries unique context that is essential for accurate detection. A global effort is needed to build datasets that cover many languages and include various media formats. Using techniques like transfer learning and cross-lingual training can help models generalize better across different languages and regions. Solving these issues is essential for creating a comprehensive defense against misinformation. Future research can prioritize the creation of benchmark datasets that include a wide range of modalities and languages. Additionally, improving how models detect contradictions between text and visuals will provide better insights into the authenticity of news content. By focusing on these areas, researchers can move toward more reliable and globally applicable systems for detecting fake news [178].

6.2.4 Scalability, Adaptability and Others

Fake news detection systems face several obstacles regarding scalability and adaptability, which are related to dataset and multimodal approaches mentioned above. Scalability is difficult because social media platforms generate a large volume of data that requires processing in real time. Many existing models rely on heavy computational resources which makes them hard to use for high speed detection during viral events. Adaptability presents another problem as misinformation tactics change frequently. To address these issues, researchers also struggle to make a functional and scalable user platform for automatic fake news detection [179]. However, researchers still need move forward to make systems work across multiple languages and data types without losing accuracy.

The phenomenon of hallucination occurs when a machine learning model or a Large Language Model produces information that is not supported by the input data [180]. This issue makes it difficult for systems to grow because the model creates realistic but false claims that require extra work to check. For example, a model might summarize an election report

by naming a candidate who did not participate. Current research focuses on using external knowledge bases to verify facts in real time and building self-check tools to catch these errors [181]. These steps help the detection system stay accurate and adapt to different subjects without constant human monitoring.

Misuse involves the intentional application of detection tools for harmful goals such as refining propaganda. When bad actors gain access to these systems, they can test and adjust their false content until it can pass through filters without being caught. A common example is using deepfake technology to create fake videos and then using detection software to ensure the fake is hard to find. This creates a cycle where developers must constantly update their models to keep up with new threats. To counter this, current trends include the use of red-teaming methods and digital watermarks to protect information and prevent the technology from being used for harm [182].

Overgeneration happens when a detection system is too sensitive and flags real news as fake. This problem arises when models rely on simple patterns or emotional words instead of understanding the true meaning of a story. For instance, a system might incorrectly label a satirical article or a dramatic news report as misinformation. This lack of balance limits how well a system can be used in different areas like humor or opinion writing. Researchers are now developing methods that look at the context and intent of the writing to reduce these false alerts [183]. These efforts allow detection tools to handle a wider variety of content while keeping the trust of the people who use them.

Moreover, there are several challenges that fake news detection face in other area, like the lack of a universal definition for fake news, which makes it difficult to create consistent policies. Computational complexity is another problem, as finding the best nodes to block in a large social network is a difficult mathematical task. Users may also become less responsive to warnings over time, and some interventions might accidentally limit free speech or legitimate information flow. There is a survey mentioning the importance of fake news prevention and mitigation [184]. The authors review various strategies to stop the spread of false information

and reduce its impact. In the future, we can do some researches from individual, social, and government levels.

6.3 Conclusions and Future Work

Fake news detection, as an emerging and evolving research direction, has gained wide attention across various application scenarios, particularly in an era characterized by widespread digital misinformation. With the rapid development of Large Language Models, the research on AI-driven fake news detection has made significant progress in recent years. This thesis introduces a novel taxonomy that categorizes existing methodologies into two distinct lines: FND Approaches (model-centric) and FND Techniques (process-centric). We provide a holistic conceptual framework by clarifying the fake news ecosystem—including its creators, dissemination paths, and distinctive characteristics—while tracing the historical trajectory of detection from early sociological analyses to advanced machine learning and deep learning architectures. We systematically compare model methodologies across different modalities and evaluate key procedural stages of the detection pipeline, such as datasets, data augmentation, information extraction, model refinement, model validation, and results explanation. Then, the thesis provides a detailed analysis by machine learning approaches, the experiment and results showed that lighter, interpretable, and computationally efficient ML models can still achieve competitive performance benchmarks. Finally, we discuss the main challenges and trends. This thesis aims to help researchers and practitioners gain a quick overview of state-of-the-art developments in AI-driven FND and inspire future research towards building a more robust and trustworthy information ecosystem.

Future research can focus on the creation of standardized performance benchmarks to allow for direct comparisons between various fake news detection models. These evaluations should employ a consistent set of metrics like precision and the F1 score to build a clear baseline for performance. By testing models across multiple data sources, we can better

understand the generalizability of their techniques and identify which methods are effective for specific types of data. It will provide the evidence needed to guide the development of future detection systems and improve the overall clarity of the research.

Bibliography

[1] Craig Silverman. *This analysis shows how viral fake election news stories outperformed real news on Facebook*. Accessed: 2026-01-18. BuzzFeed News. Nov. 2016. URL: https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-newsoutperformed-real-news-on-facebook.

[2] Andrew S Ross and Damian J Rivers. "Discursive deflection: accusation of "fake news" and the spread of mis-and disinformation in the tweets of President Trump". In: *Social media+ society* 4.2 (2018), p. 2056305118776010.

[3] Sirisha Bojjireddy, Soon Ae Chun, and James Geller. "Machine learning approach to detect fake news, misinformation in COVID-19 pandemic". In: *Proceedings of the 22nd Annual International Conference on Digital Government Research*. dg.o '21. Omaha, NE, USA: Association for Computing Machinery, 2021, pp. 575–578. ISBN: 9781450384926.

[4] Gordon Pennycook et al. "Fighting COVID-19 misinformation on social media: experimental evidence for a scalable accuracy-nudge intervention". In: *Psychological science* 31.7 (2020), pp. 770–780.

[5] Shimon Kogan, Tobias J Moskowitz, and Marina Niessner. "Fake news: evidence from financial markets". In: *Available at SSRN 3231461* (2018).

[6] Xiangji Huang et al. “Applying machine learning to text segmentation for information retrieval”. In: *Inf. Retr.* 6.3-4 (2003), pp. 333–362.

[7] Zhaohui Liang et al. “Deep learning for healthcare decision making with EMRs”. In: *2014 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2014, Belfast, United Kingdom, November 2-5, 2014*. Ed. by Huiru Jane Zheng et al. IEEE Computer Society, 2014, pp. 556–559.

[8] Runjie Zhu, Xinhui Tu, and Jimmy Xiangji Huang. “Deep learning on information retrieval and its applications”. In: *Deep learning for data analytics*. Elsevier, 2020, pp. 125–153.

[9] Israt Jahan et al. “Evaluation of ChatGPT on biomedical tasks: a zero-shot comparison with Fine-Tuned Generative Transformers”. In: *CoRR* abs/2306.04504 (2023). arXiv: 2306.04504.

[10] Yizheng Huang and Jimmy X. Huang. “Exploring ChatGPT for next-generation information retrieval: opportunities and challenges”. In: *Web Intell.* 22.1 (2024), pp. 31–44.

[11] Kai Shu et al. “Fake news detection on social media: A data mining perspective”. In: *ACM SIGKDD explorations newsletter* 19.1 (2017), pp. 22–36.

[12] Xinyi Zhou and Reza Zafarani. “A survey of fake news: fundamental theories, detection methods, and opportunities”. In: *ACM Computing Surveys (CSUR)* 53.5 (2020), pp. 1–40.

[13] Yasmim Mendes Rocha et al. “The impact of fake news on social media and its influence on health during the COVID-19 pandemic: a systematic review”. In: *Journal of Public Health* (2021), pp. 1–10.

[14] Alexandre Bovet and Hernán A Makse. “Influence of fake news in Twitter during the 2016 US presidential election”. In: *Nature communications* 10.1 (2019), p. 7.

[15] Tanja Pavleska et al. “Performance analysis of fact-checking organizations and initiatives in Europe: a critical overview of online platforms fighting fake news”. In: *Social media and convergence* 29 (2018), pp. 1–28.

[16] Kelley Cotter, Julia R DeCook, and Shaheen Kanthawala. “Fact-checking the crisis: COVID-19, infodemics, and the platformization of truth”. In: *Social Media+ Society* 8.1 (2022), p. 20563051211069048.

[17] Ángel Vizoso and Jorge Vázquez-Herrero. “Fact-checking platforms in Spanish. Features, organisation and method”. In: *Communication & society* (2019), pp. 127–142.

[18] SY Yuliani et al. “A framework for hoax news detection and analyzer used rule-based methods”. In: *International Journal of Advanced Computer Science and Applications* 10.10 (2019).

[19] Xiangji Huang et al. “Applying data mining to pseudo-relevance feedback for high performance text retrieval”. In: *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM 2006), 18-22 December 2006, Hong Kong, China*. IEEE Computer Society, 2006, pp. 295–306.

[20] Supanya Aphiwongsophon and Prabhas Chongstitvatana. “Detecting fake news with machine learning method”. In: *2018 15th international conference on electrical engineering/electronics, computer, telecommunications and information technology (ECTI-CON)*. IEEE. 2018, pp. 528–531.

[21] Seyedmehdi Hosseinimotlagh and Evangelos E Papalexakis. “Unsupervised content-based identification of fake news articles with tensor decomposition ensembles”. In: *Proceedings of the Workshop on Misinformation and Misbehavior Mining on the Web (MIS2)*. 2018.

[22] Muhammad Firoz Mridha et al. “A comprehensive review on fake news detection with deep learning”. In: *IEEE access* 9 (2021), pp. 156151–156170.

[23] K Anirudh, Meghana Srikanth, and A Shahina. "Multilingual fake news detection in low-resource languages: a comparative study using BERT and GPT-3.5". In: *International Conference on Speech and Language Technologies for Low-resource Languages*. Springer. 2023, pp. 387–397.

[24] Sabrine Amri, Dorsaf Sallami, and Esma Aïmeur. "EXMULF: an explainable multi-modal content-based fake news detection system". In: *Foundations and Practice of Security*. Ed. by Esma Aïmeur et al. Cham: Springer International Publishing, 2022, pp. 177–187.

[25] Jing Shen et al. "Multi-modal similarity guided adaptive fusion network for short video fake news detection". In: *Proceedings of the 2025 International Conference on Multimedia Retrieval*. 2025, pp. 1145–1153.

[26] Peng Qi et al. "Sniffer: multimodal large language model for explainable out-of-context misinformation detection". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024, pp. 13052–13062.

[27] Ala Mughaid et al. "An intelligent cybersecurity system for detecting fake news in social media websites". In: *Soft Computing* 26.12 (2022), pp. 5577–5591.

[28] Kristína Machová, Marián Mach, and Viliam Balara. "Federated learning in the detection of fake news using deep learning as a basic method". In: *Sensors* 24.11 (2024), p. 3590.

[29] Rafał Kozik et al. "When explainability turns into a threat-using xAI to fool a fake news detection method". In: *Computers & Security* 137 (2024), p. 103599.

[30] Sabrine Amri, Henri–Cedric Mputu Boleilanga, and Esma Aimeur. "ExFake: towards an explainable fake news detection based on content and social context information". In: *Proceedings of 2023 Congress in Computer Science, Computer Engineering, Applied Computing (CSCE)*. 2023, pp. 01–08.

[31] Hunt Allcott and Matthew Gentzkow. “Social media and fake news in the 2016 election”. In: *Journal of Economic Perspectives* 31.2 (2017), pp. 211–236.

[32] Arianna D’ulizia et al. “Fake news detection: a survey of evaluation datasets”. In: *PeerJ Computer Science* 7 (2021), e518.

[33] Esma Aımeur, Sabrine Amri, and Gilles Brassard. “Fake news, disinformation and misinformation in social media: a review”. In: *Social Network Analysis and Mining* 13.1 (2023), p. 30.

[34] Shivani Tufchi, Ashima Yadav, and Tanveer Ahmed. “A comprehensive survey of multimodal fake news detection techniques: advances, challenges, and opportunities”. In: *International Journal of Multimedia Information Retrieval* 12.2 (2023), p. 28.

[35] Carmela Comito, Luciano Caroprese, and Ester Zumpano. “Multimodal fake news detection on social media: a survey of deep learning techniques”. In: *Social Network Analysis and Mining* 13.1 (2023), p. 101.

[36] Xianghua Li et al. “A survey of multimodal fake news detection: a cross-modal interaction perspective”. In: *IEEE Transactions on Emerging Topics in Computational Intelligence* (2025).

[37] Xichen Zhang and Ali A Ghorbani. “An overview of online fake news: characterization, detection, and discussion”. In: *Information Processing & Management* 57.2 (2020), p. 102025.

[38] Eugène Loos and Jordy Nijenhuis. “Consuming fake news: a matter of age? the perception of political fake news stories in Facebook ads”. In: *Human Aspects of IT for the Aged Population. Technology and Society: 6th International Conference, ITAP 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part III 22*. Springer. 2020, pp. 69–88.

[39] Sander Van der Linden, Costas Panagopoulos, and Jon Roozenbeek. "You are fake news: political bias in perceptions of fake news". In: *Media, culture & society* 42.3 (2020), pp. 460–470.

[40] Ann Devitt and Khurshid Ahmad. "Sentiment polarity identification in financial news: a cohesion-based approach". In: *Proceedings of the 45th annual meeting of the association of computational linguistics*. 2007, pp. 984–991.

[41] John P Dickerson, Vadim Kagan, and VS Subrahmanian. "Using sentiment to detect bots on twitter: are humans more opinionated than bots?" In: *2014 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM 2014)*. IEEE. 2014, pp. 620–627.

[42] Victoria L Rubin, Niall J Conroy, and Yimin Chen. "Towards news verification: deception detection methods for news discourse". In: *Hawaii international conference on system sciences*. 2015, pp. 5–8.

[43] Adrian MP Braşoveanu and Răzvan Andonie. "Semantic fake news detection: a machine learning perspective". In: *International Work-Conference on Artificial Neural Networks*. Springer. 2019, pp. 656–667.

[44] Jia Wang et al. "Find: fine-grained discrepancy-based fake news detection enhanced by event abstract generation". In: *Computer Speech & Language* 78 (2023), p. 101461.

[45] Bobby Chesney and Danielle Citron. "Deep fakes: a looming challenge for privacy, democracy, and national security". In: *Calif. L. Rev.* 107 (2019), p. 1753.

[46] Jing Jing et al. "Multimodal fake news detection via progressive fusion networks". In: *Information processing & management* 60.1 (2023), p. 103120.

[47] Soroush Vosoughi, Deb Roy, and Sinan Aral. "The spread of true and false news online". In: *science* 359.6380 (2018), pp. 1146–1151.

[48] Mona Nasery. “Fake news on social media: from fake news lifecycle to fake news combat cycle”. PhD thesis. 2024.

[49] Simone Raponi et al. “Fake news propagation: a review of epidemic models, datasets, and insights”. In: *ACM Transactions on the Web (TWEB)* 16.3 (2022), pp. 1–34.

[50] Jennifer Lackey. “Echo chambers, fake news, and social epistemology”. In: *The epistemology of fake news* (2021), pp. 206–227.

[51] Botambu Collins et al. “Trends in combating fake news on social media–a survey”. In: *Journal of Information and Telecommunication* 5.2 (2021), pp. 247–266.

[52] Mykhailo Granik and Volodymyr Mesyura. “Fake news detection using naive bayes classifier”. In: *2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON)*. IEEE. 2017, pp. 900–903.

[53] Nihel Fatima Baarir and Abdelhamid Djeffal. “Fake news detection using machine learning”. In: *2020 2nd International workshop on human-centric smart environments for health and well-being (IHSH)*. IEEE. 2021, pp. 125–130.

[54] N Leela Siva Rama Krishna and M Adimoolam. “Fake news detection system using decision tree algorithm and compare textual property with support vector machine algorithm”. In: *2022 International conference on business analytics for technology and security (ICBATS)*. IEEE. 2022, pp. 1–6.

[55] Zeba Khanam et al. “Fake news detection using machine learning approaches”. In: *Proceedings of IOP conference series: materials science and engineering*. Vol. 1099. 1. IOP Publishing. 2021, p. 012040.

[56] Iftikhar Ahmad et al. “Fake news detection using machine learning ensemble methods”. In: *Complexity* 2020.1 (2020), p. 8885861.

[57] Ankit Kesarwani, Sudakar Singh Chauhan, and Anil Ramachandran Nair. "Fake news detection on social media using k-nearest neighbor classifier". In: *2020 international conference on advances in computing and communication engineering (ICACCE)*. IEEE. 2020, pp. 1–4.

[58] Muhammad Umer et al. "Fake news stance detection using deep learning architecture (CNN-LSTM)". In: *IEEE Access* 8 (2020), pp. 156695–156706.

[59] Qian Li et al. "Multi-level word features based on CNN for fake news detection in cultural communication". In: *Personal and Ubiquitous Computing* 24 (2020), pp. 259–272.

[60] Pritika Bahad, Preeti Saxena, and Raj Kamal. "Fake news detection using bi-directional LSTM-recurrent neural network". In: *Procedia Computer Science* 165 (2019), pp. 74–82.

[61] Xu Chen et al. "Neural feature-aware recommendation with signed hypergraph convolutional network". In: *ACM Transactions on Information Systems (TOIS)* 39.1 (2020), pp. 1–22.

[62] Jamal Abdul Nasir, Osama Subhani Khan, and Iraklis Varlamis. "Fake news detection: a hybrid CNN-RNN based deep learning approach". In: *International journal of information management data insights* 1.1 (2021), p. 100007.

[63] Qin Chen et al. "CA-RNN: using context-aligned recurrent neural networks for modeling sentence similarity". In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), 2018*. Ed. by Sheila A. McIlraith and Kilian Q. Weinberger. AAAI Press, 2018, pp. 265–273.

[64] Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach". In: *Multimedia tools and applications* 80.8 (2021), pp. 11765–11788.

[65] Mateusz Szczepański et al. “New explainability method for BERT-based model in fake news detection”. In: *Scientific reports* 11.1 (2021), p. 23705.

[66] Mohit Mayank, Shakshi Sharma, and Rajesh Sharma. “DEAP-FAKED: Knowledge graph based approach for fake news detection”. In: *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE. 2022, pp. 47–51.

[67] Shuzhi Gong et al. “Fake news detection through graph-based neural networks: a survey”. In: *arXiv preprint arXiv:2307.12639* (2023).

[68] Karine Aoun Barakat, Amal Dabbous, and Abbas Tarhini. “An empirical approach to understanding users’ fake news identification on social media”. In: *Online Information Review* 45.6 (2021), pp. 1080–1096.

[69] Yingtong Dou et al. “User preference-aware fake news detection”. In: *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 2021, pp. 2051–2055.

[70] Nirosh Jayakody, Azeem Mohammad, and Malka N Halgamuge. “Fake news detection using a decentralized deep learning model and federated learning”. In: *IECON 2022–48th Annual Conference of the IEEE Industrial Electronics Society*. IEEE. 2022, pp. 1–6.

[71] Wenhao Li, Selvakumar Manickam, and Yung-Wey Chong. “FedPhishLLM: a privacy-preserving and explainable phishing detection mechanism using federated learning and LLMs”. In: *Journal of King Saud University Computer and Information Sciences* 37.8 (2025), p. 252.

[72] Aijun An et al. “Feature selection with rough sets for web page classification”. In: *Transactions on Rough Sets II: Rough Sets and Fuzzy Sets*. Springer, 2004, pp. 1–13.

[73] Fuchun Peng et al. “Investigating the relationship between word segmentation performance and retrieval performance in Chinese IR”. In: *COLING 2002: The 19th International Conference on Computational Linguistics*. 2002.

[74] Md Tahmid Rahman Laskar, Enamul Hoque, and Xiangji Huang. “WSL-DS: weakly supervised learning with distant supervision for query focused multi-document abstractive summarization”. In: *Proceedings of the 28th International Conference on Computational Linguistics*. 2020, pp. 5647–5654.

[75] Tom Fawcett. “An introduction to ROC analysis”. In: *Pattern Recognition Letters* 27.8 (2006). ROC Analysis in Pattern Recognition, pp. 861–874. ISSN: 0167-8655.

[76] Davide Chicco and Giuseppe Jurman. “The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification”. In: *BioData Mining* 16.1 (2023), p. 4.

[77] Liu Yang et al. “Gpteval: Nlg evaluation using gpt-4 with better human alignment”. In: *arXiv preprint arXiv: 2303.16634* (2023).

[78] Kyungha Kim et al. “Can llms produce faithful explanations for fact-checking? towards faithful explainable fact-checking via multi-agent debate”. In: *arXiv preprint arXiv:2402.07401* (2024).

[79] Quanzhi Li et al. “Rumor detection on social media: Datasets, methods and opportunities”. In: *arXiv preprint arXiv:1911.07199* (2019).

[80] James Thorne et al. “FEVER: a large-scale dataset for fact extraction and verification”. In: *arXiv preprint arXiv:1803.05355* (2018).

[81] Xiaomo Liu et al. “Real-time rumor debunking on Twitter”. In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. CIKM ’15. Melbourne, Australia: Association for Computing Machinery, 2015, pp. 1867–1870. ISBN: 9781450337946.

[82] Jing Ma et al. “Detecting rumors from microblogs with recurrent neural networks”. In: *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*. New York, NY, USA: AAAI Press, 2016, pp. 3818–3824. URL: https://ink.library.smu.edu.sg/sis_research/4630.

[83] Kai Shu et al. “FakeNewsNet: a data repository with news content, social context, and spatiotemporal information for studying fake news on social media”. In: *Big data* 8.3 (2020), pp. 171–188.

[84] Tsun-Hin Cheung and Kin-Man Lam. “Factllama: Optimizing instruction-following language models with external knowledge for automated fact-checking”. In: *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE. 2023, pp. 846–853.

[85] Liangming Pan et al. “Fact-checking complex claims with program-guided reasoning”. In: *arXiv preprint arXiv:2305.12744* (2023).

[86] Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. “Large language model agent for fake news detection”. In: *arXiv preprint arXiv:2405.01593* (2024).

[87] Arkaitz Zubiaga et al. “Analysing how people orient to and spread rumours in social media by looking at conversational threads”. In: *PLOS ONE* 11 (Mar. 2016), pp. 1–29.

[88] C. Silverman et al. *Hyperpartisan Facebook pages are publishing false and misleading information at an alarming rate*. Accessed: 2026-01-18. BuzzFeed News. Oct. 2016. URL: https://www.buzzfeed.com/craigsilverman/partisan-fb-pages-analysis.

[89] William Yang Wang. ““liar, liar pants on fire”: a new benchmark dataset for fake news detection”. In: *arXiv preprint arXiv:1705.00648* (2017).

[90] Abderrazek Azri et al. “DAT@ Z21: a comprehensive multimodal dataset for rumor classification in microblogs”. In: *International Conference on Big Data Analytics and Knowledge Discovery*. Springer. 2023, pp. 161–175.

[91] Yang He et al. “LTCR: long temporal characteristic reconstruction for segmentation in contrastive learning”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2024, pp. 355–371.

[92] Megha Sundriyal, Tanmoy Chakraborty, and Preslav Nakov. “From chaos to clarity: claim normalization to empower fact-checking”. In: *arXiv preprint arXiv:2310.14338* (2023).

[93] Lionel Z Wang et al. “Megafake: a theory-driven dataset of fake news generated by large language models”. In: *arXiv preprint arXiv:2408.11871* (2024).

[94] Jason Lucas et al. “Fighting fire with fire: The dual role of LLMs in crafting and detecting elusive disinformation”. In: *arXiv preprint arXiv:2310.15515* (2023).

[95] Yue Huang and Lichao Sun. “FakeGPT: fake news generation, explanation and detection of large language models”. In: *arXiv preprint arXiv:2310.05046* (2023).

[96] Kung-Hsiang Huang et al. “Faking fake news for real fake news detection: Propaganda-loaded training data generation”. In: *arXiv preprint arXiv:2203.05386* (2022).

[97] Beizhe Hu et al. “Bad actor, good advisor: exploring the role of large language models in fake news detection”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 20. 2024, pp. 22105–22113.

[98] Zizhong Li, Haopeng Zhang, and Jiawei Zhang. “A revisit of fake news dataset with augmented fact-checking by chatgpt”. In: *arXiv preprint arXiv:2312.11870* (2023).

[99] Yupeng Cao et al. “Can large language models detect misinformation in scientific news reporting?” In: *arXiv preprint arXiv:2402.14268* (2024).

[100] Longzheng Wang et al. “Mmidr: teaching large language model to interpret multimodal misinformation via knowledge distillation”. In: *arXiv preprint arXiv:2403.14171* (2024).

[101] Chang Yang et al. “Rumor detection on social media with crowd intelligence and ChatGPT-assisted networks”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2023, pp. 5705–5717.

[102] Marian Bucos and Georgiana Ţucudean. “Text data augmentation techniques for fake news detection in the Romanian language”. In: *Applied Sciences* 13.13 (2023), p. 7389.

[103] Jozef Kapusta et al. “Text data augmentation techniques for word embeddings in fake news classification”. In: *IEEE Access* 12 (2024), pp. 31538–31550.

[104] Abdelhalim Hafedh Dahou et al. “Enhancing model performance through translation-based data augmentation in the context of fake news detection”. In: *Procedia Computer Science* 244 (2024), pp. 342–352.

[105] Jiaying Wu, Jiafeng Guo, and Bryan Hooi. “Fake news in sheep’s clothing: robust fake news detection against LLM-empowered style attacks”. In: *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*. 2024, pp. 3367–3378.

[106] Yike Wu et al. “Towards robust evidence-aware fake news detection via improving semantic perception”. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 2024, pp. 16607–16618.

[107] Ke Wang et al. “A survey on data synthesis and augmentation for large language models”. In: *arXiv preprint arXiv:2410.12896* (2024).

[108] Fei Wu et al. “Efficient cross-modal prompt learning with semantic enhancement for domain-robust fake news detection”. In: *Proceedings of the 31st International Conference on Computational Linguistics*. 2025, pp. 4175–4185.

[109] Ye Jiang et al. “Cross-modal augmentation for few-shot multimodal fake news detection”. In: *Engineering Applications of Artificial Intelligence* 142 (2025), p. 109931.

[110] Longzheng Wang et al. “Cross-modal contrastive learning for multimodal fake news detection”. In: *Proceedings of the 31st ACM international conference on multimedia*. 2023, pp. 5696–5704.

[111] Maaz Amjad, Grigori Sidorov, and Alisa Zhila. “Data augmentation using machine translation for fake news detection in the Urdu language”. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. 2020, pp. 2537–2542.

[112] Jianqiao Lai et al. “Rumorllm: A rumor large language model-based fake-news-detection data-augmentation approach”. In: *Applied Sciences* 14.8 (2024), p. 3532.

[113] Herun Wan et al. “Dell: Generating reactions and explanations for LLM-based misinformation detection”. In: *arXiv preprint arXiv:2402.10426* (2024).

[114] Qiong Nan et al. “Let silence speak: enhancing fake news detection with generated comments from large language models”. In: *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2024, pp. 1732–1742.

[115] Shan Jia et al. “AutoSplice: a text-prompt manipulated image dataset for media forensics”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 893–903.

[116] Qinghao Ye et al. “mPLUG-Owl2: revolutionizing multi-modal large language model with modality collaboration”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024, pp. 13040–13051.

[117] Mohammad Majid Akhtar et al. “Towards automatic annotation and detection of fake news”. In: *Proceedings of 2023 IEEE 48th Conference on Local Computer Networks (LCN)*. 2023, pp. 1–9.

[118] Madina Sambetbayeva et al. “A multi-level annotation model for fake news detection: implementing Kazakh-Russian corpus via Label Studio”. In: *Big Data and Cognitive Computing* 9.8 (2025), p. 215.

[119] Miaoran Li et al. “Self-checker: Plug-and-play modules for fact-checking with large language models”. In: *arXiv preprint arXiv:2305.14623* (2023).

[120] Nicola Capuano et al. “Content-based fake news detection with machine and deep learning: a systematic review”. In: *Neurocomputing* 530 (2023), pp. 91–103.

[121] Mirmorsal Madani, Homayun Motameni, and Hosein Mohamadi. “Fake news detection using deep learning integrating feature extraction, natural language processing, and statistical descriptors”. In: *Security and Privacy* 5.6 (2022), e264.

[122] Yuhang Wang et al. “Detecting fake news by enhanced text representation with multi-EDU-structure awareness”. In: *Expert Systems with Applications* 206 (2022), p. 117781.

[123] Yanfang Qiu et al. “DSEN-EK: dual-layer semantic information extraction network with external knowledge for fake news detection”. In: *International Journal of Web Information Systems* 21.2 (2025), pp. 139–157.

[124] Peng Qi et al. “Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues”. In: *Proceedings of the 29th ACM international conference on multimedia*. 2021, pp. 1212–1220.

[125] Guangyang Wu et al. “Cheap-fake detection with LLM using prompt engineering”. In: *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. IEEE. 2023, pp. 105–109.

[126] Bohdan M Pavlyshenko. “Analysis of disinformation and fake news detection using fine-tuned large language model”. In: *arXiv preprint arXiv:2309.04704* (2023).

[127] Xin Tan, Bowei Zou, and Ai Ti Aw. “Evidence-based interpretable open-domain fact-checking with large language models”. In: *arXiv preprint arXiv:2312.05834* (2023).

[128] Dorian Quelle and Alexandre Bovet. “The perils and promises of fact-checking with large language models”. In: *Frontiers in Artificial Intelligence* 7 (2024), p. 1341697.

[129] I Chern et al. “FacTool: factuality detection in generative AI–a tool augmented framework for multi-task and multi-domain scenarios”. In: *arXiv preprint arXiv:2307.13528* (2023).

[130] Yizheng Huang and Jimmy Huang. “A survey on retrieval-augmented text generation for large language models”. In: *arXiv preprint arXiv:2404.10981* (2024).

[131] Yizheng Huang and Jimmy X. Huang. “Diversified prior knowledge enhanced general language model for biomedical information retrieval”. In: *ECAI 2023 - 26th European Conference on Artificial Intelligence, September 30 - October 4, 2023, Kraków, Poland - Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023)*. Ed. by Kobi Gal et al. Vol. 372. Frontiers in Artificial Intelligence and Applications. IOS Press, 2023, pp. 1109–1115.

[132] Yizheng Huang and Jimmy X. Huang. “York University at TREC 2021: deep learning track”. In: *Proceedings of the Thirtieth Text REtrieval Conference, TREC 2021, online, November 15-19, 2021*. Ed. by Ian Soboroff and Angela Ellis. Vol. 500-335. NIST Special Publication. National Institute of Standards and Technology (NIST), 2021.

[133] Patrick Lewis et al. “Retrieval-augmented generation for knowledge-intensive NLP tasks”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 9459–9474.

[134] Jack W Rae et al. “Scaling language models: methods, analysis & insights from training Gopher”. In: *arXiv preprint arXiv:2112.11446* (2021).

[135] Ronit Singal et al. “Evidence-backed fact checking using RAG and few-shot in-context learning with LLMs”. In: *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*. 2024, pp. 91–98.

[136] Cheng Niu et al. “VeraCT scan: retrieval-augmented fake news detection with justifiable reasoning”. In: *arXiv preprint arXiv:2406.10289* (2024).

[137] Guanghua Li et al. “Re-search for the truth: multi-round retrieval-augmented large language models are strong fake news detectors”. In: *arXiv preprint arXiv:2403.09747* (2024).

[138] Mohammed Abdul Khaliq et al. “RAGAR, your falsehood radar: RAG-augmented reasoning for political fact-checking using multimodal large language models”. In: *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*. 2024, pp. 280–296.

[139] Yangxiao Bai and Kaiqun Fu. “A large language model-based fake news detection framework with RAG fact-checking”. In: *Proceedings of 2024 IEEE International Conference on Big Data (BigData)*. IEEE. 2024, pp. 8617–8619.

[140] Sonali Singh and Akbar Siami Namin. “Adversarial training of retrieval augmented generation to generate believable fake news”. In: *Proceedings of 2024 IEEE International Conference on Big Data (BigData)*. IEEE. 2024, pp. 3589–3598.

[141] Linda Zeng et al. “Worse than zero-shot? a fact-checking dataset for evaluating the robustness of RAG against misleading retrievals”. In: *arXiv preprint arXiv:2502.16101* (2025).

[142] Mohammad Vatani Nezafat and Saeed Samet. “Fake news detection with retrieval augmented generative artificial intelligence”. In: *Proceedings of 2024 2nd International Conference on Foundation and Large Language Models (FLLM)*. IEEE. 2024, pp. 160–167.

[143] Hai Li et al. “Use of retrieval-augmented large language model for COVID-19 fact-checking: development and usability study”. In: *Journal of Medical Internet Research* 27 (2025), e66098.

[144] Rishabh Upadhyay and Marco Viviani. "Enhancing health information retrieval with RAG by prioritizing topical relevance and factual accuracy". In: *Discover Computing* 28.1 (2025), p. 27.

[145] Akari Asai et al. "Self-RAG: learning to retrieve, generate, and critique through self-reflection". In: *arXiv preprint arXiv:2310.11511* (2023).

[146] Zhiwei Liu et al. "RaemoLLM: retrieval augmented LLMs for cross-domain misinformation detection using in-context learning based on emotional information". In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2025, pp. 16508–16523.

[147] Zhihong Shao et al. "Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy". In: *arXiv preprint arXiv:2305.15294* (2023).

[148] Ye Liu et al. "Detect, investigate, judge and determine: a novel llm-based framework for few-shot fake news detection". In: *arXiv preprint arXiv:2407.08952* (2024).

[149] Erik Cambria et al. "Xai meets llms: a survey of the relation between explainable ai and large language models". In: *arXiv preprint arXiv:2407.15248* (2024).

[150] Xuansheng Wu et al. "Usable XAI: 10 strategies towards exploiting explainability in the LLM era". In: *arXiv preprint arXiv:2403.08946* (2024).

[151] Haoran Wang and Kai Shu. "Explainable claim verification via knowledge-grounded reasoning with large language models". In: *arXiv preprint arXiv:2310.05253* (2023).

[152] Xuan Zhang and Wei Gao. "Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method". In: *arXiv preprint arXiv:2310.00305* (2023).

[153] Ganqu Cui et al. "Ultrafeedback: boosting language models with scaled AI feedback". In: *arXiv preprint arXiv:2310.01377* (2023).

[154] Tanushree Banerjee et al. “LLMs are superior feedback providers: bootstrapping reasoning for lie detection with self-generated feedback”. In: *arXiv preprint arXiv:2408.13915* (2024).

[155] Kai Shu et al. “Fake news detection on social media: a data mining perspective”. In: *ACM SIGKDD Explorations Newsletter* 19 (1 June 2017), pp. 22–36.

[156] Kai Shu, Suhang Wang, and Huan Liu. “Exploiting tri-relationship for fake news detection”. In: *arXiv preprint arXiv:1712.07709* (Dec. 2017).

[157] Li Zeng and Xiaoci Tao. “Machine learning-based fake news detection on social media”. In: *2024 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. IEEE. 2024, pp. 687–692.

[158] Kai Shu et al. “FakeNewsNet: a data repository with news content, social context, and dynamic information for studying fake news on social media”. In: *arXiv preprint arXiv:1809.01286* (Sept. 2018).

[159] P. Meel and D. K. Vishwakarma. “A temporal ensembling based semi-supervised ConvNet for the detection of fake news articles”. In: *Expert Systems with Applications* 177 (Oct. 2021), p. 115002.

[160] L. I. Meng, L. I. Yanling, and L. I. N. Min. “Review of transfer learning for named entity recognition”. In: *Journal of Frontiers of Computer Science Technology* 15 (2 2021).

[161] C. M. Tsai. “Stylometric fake news detection based on natural language processing using named entity recognition: in-domain and cross-domain analysis”. In: *Electronics* 12 (17 2023), p. 3676.

[162] Nagarajan Ganapathy, Yedukondala Rao Veeranki, and Ramakrishnan Swaminathan. “Convolutional neural network based emotion classification using electrodermal activity

signals and time-frequency features". In: *Expert Systems with Applications* 159 (2020), p. 113571.

[163] Mykhailo Granik and Volodymyr Mesyura. "Fake news detection using naive bayes classifier". In: *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*. 2017, pp. 900–903. DOI: 10.1109/UKRCON.2017.8100379.

[164] Reham Jehad and Suhad A.Yousif. "Fake news classification using random forest and decision tree (J48)". In: *Al-Nahrain Journal of Science* (2020). URL: https://api.semanticscholar.org/CorpusID:229394693.

[165] Joseph Meynard Ogdol and Bill-Lawrence Samar. "Binary logistic regression based classifier for fake news". In: (June 2018).

[166] Fatemeh Torabi Asr and Maite Taboada. "Big data and quality data for fake news and misinformation detection". In: *Big data & society* 6.1 (2019), p. 2053951719843310.

[167] Soveatin Kuntur et al. "Fake news detection: it's all in the data!" In: *arXiv preprint arXiv:2407.02122* (2024).

[168] Ana Đurić et al. "Assessing reproducibility and accessibility in hate speech and fake news detection datasets: a literature review (2023-2024)". In: *Proceedings of 2025 24th International Symposium INFOTEH-JAHORINA (INFOTEH)*. IEEE. 2025, pp. 1–6.

[169] Suhaib Kh Hamed, Mohd Juzaiddin Ab Aziz, and Mohd Ridzwan Yaakub. "A review of fake news detection approaches: a critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion". In: *Heliyon* 9.10 (2023).

[170] Kenneth M Merz Jr et al. *Method and data sharing and reproducibility of scientific results*. 2020.

[171] Marjan Hosseini et al. "Interpretable fake news detection with topic and deep variational models". In: *Online Social Networks and Media* 36 (2023), p. 100249.

[172] Zheng Ye, Jimmy Xiangji Huang, and Hongfei Lin. "Finding a good query-related topic for boosting pseudo-relevance feedback". In: *J. Assoc. Inf. Sci. Technol.* 62.4 (2011), pp. 748–760.

[173] Benjamin D Horne, Dorit Nevo, and Susan L Smith. "Ethical and safety considerations in automated fake news detection". In: *Behaviour & Information Technology* (2023), pp. 1–22.

[174] Sina Mohseni, Eric Ragan, and Xia Hu. "Open issues in combating fake news: interpretability as an opportunity". In: *arXiv preprint arXiv:1904.03016* (2019).

[175] Junxiao Xue et al. "Detecting fake news by exploring the consistency of multimodal data". In: *Information Processing & Management* 58.5 (2021), p. 102610.

[176] Shubha Mishra, Piyush Shukla, and Ratish Agarwal. "Analyzing machine learning enabled fake news detection techniques for diversified datasets". In: *Wireless Communications and Mobile Computing* 2022.1 (2022), p. 1575365.

[177] Rami Mohawesh, Sumbal Maqsood, and Qutaibah Althebyan. "Multilingual deep learning framework for fake news detection using capsule neural network". In: *Journal of Intelligent Information Systems* 60.3 (2023), pp. 655–671.

[178] Arkaitz Zubiaga et al. "Detection and resolution of rumours in social media: a survey". In: *Acm Computing Surveys (Csur)* 51.2 (2018), pp. 1–36.

[179] Sari Stissi. "A functional and scale-able user platform for automatic fake news detection". In: *Semantic Scholar* (2020).

[180] Ziwei Ji et al. "Survey of hallucination in natural language generation". In: *ACM Computing Surveys* 55.12 (2023), pp. 1–38.

[181] Lei Huang et al. "A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions". In: *ACM Transactions on Information Systems* 43.2 (2025), pp. 1–55.

[182] David Megías et al. “Architecture of a fake news detection system combining digital watermarking, signal processing, and machine learning”. In: *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), 2022, 13 (1): 33-55,* (2022).

[183] Anuj Kumar et al. “KGFakeNet: a knowledge graph-enhanced model for fake news detection”. In: *Proceedings of the Workshop on Generative AI and Knowledge Graphs (GenAIK)*. 2025, pp. 109–122.

[184] Dorsaf Sallami and Esma Aımeur. “Exploring beyond detection: a review on fake news prevention and mitigation techniques”. In: *Journal of Computational Social Science* 8.1 (2025), p. 23.

Appendix A

Published Papers and Papers Under Review

1. Yizheng Huang and Li Zeng. "Multiple Linear Combination Approaches for Information Search in Ranking". In: *2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. IEEE, 2022. Published.

2. Li Zeng and Xiaoci Tao. "Machine Learning-Based Fake News Detection on Social Media". In: 2024 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT). 2024. Published. The Best Student Paper Award.

3. Li Zeng, Yizheng Huang and Jimmy Huang. "Fighting Fake News: A Survey of AI-based Approaches, Techniques and Challenges". In: *Journal, ACM Computing Surveys*. 2025. Conditionally accepted with revision, under 2ed round review.

4. Li Zeng, Ellen Huang and Yizheng Huang. "The Performance Evaluation of Large Language Models in Assisting the Diagnosis and Treatment of Hashimoto's Thyroiditis ". In: *Journal, Natural Language Processing Journal*. 2025. Under review.