future internet

MDPI

Review

A Survey on the Use of Large Language Models (LLMs) in Fake News

Eleftheria Papageorgiou [1], Christos Chronis [1], Iraklis Varlamis [1,*] and Yassine Himeur [2]

[1] Department of Informatics and Telematics, Harokopio University of Athens, GR-17778 Athens, Greece; csi23310@hua.gr (E.P.); chronis@hua.gr (C.C.)
[2] College of Engineering and Information Technology, University of Dubai, Dubai P.O. Box 14143, United Arab Emirates; yhimeur@ud.ac.ae
* Correspondence: varlamis@hua.gr

Abstract: The proliferation of fake news and fake profiles on social media platforms poses significant threats to information integrity and societal trust. Traditional detection methods, including rule-based approaches, metadata analysis, and human fact-checking, have been employed to combat disinformation, but these methods often fall short in the face of increasingly sophisticated fake content. This review article explores the emerging role of Large Language Models (LLMs) in enhancing the detection of fake news and fake profiles. We provide a comprehensive overview of the nature and spread of disinformation, followed by an examination of existing detection methodologies. The article delves into the capabilities of LLMs in generating both fake news and fake profiles, highlighting their dual role as both a tool for disinformation and a powerful means of detection. We discuss the various applications of LLMs in text classification, fact-checking, verification, and contextual analysis, demonstrating how these models surpass traditional methods in accuracy and efficiency. Additionally, the article covers LLM-based detection of fake profiles through profile attribute analysis, network analysis, and behavior pattern recognition. Through comparative analysis, we showcase the advantages of LLMs over conventional techniques and present case studies that illustrate practical applications. Despite their potential, LLMs face challenges such as computational demands and ethical concerns, which we discuss in more detail. The review concludes with future directions for research and development in LLM-based fake news and fake profile detection, underscoring the importance of continued innovation to safeguard the authenticity of online information.

Keywords: fake news; fake profiles; fact-checking; large language models (LLMs); text classification

check for updates

Citation: Papageorgiou, E.; Chronis, C.; Varlamis, I.; Himeur, Y. A Survey on the Use of Large Language Models (LLMs) in Fake News. *Future Internet* **2024**, *16*, 298. https://doi.org/10.3390/fi16080298

Academic Editor: Filipe Portela

Received: 29 July 2024
Revised: 12 August 2024
Accepted: 17 August 2024
Published: 19 August 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, the proliferation of fake news has become a significant concern in the digital age. With the rise of social media platforms and the increasing ease of information dissemination, the spread of misinformation and disinformation poses serious threats to public opinion, democratic processes, and societal harmony. Fake news and fake profiles, often used in conjunction with sophisticated astroturfing techniques, manipulate public perception and can lead to widespread misinformation [1].

Several deep learning models have been recommended in the recent past to answer the challenges of detecting fake news and their dissemination patterns, such as Long-short term memory (LSTM) and Convolutional Neural Networks (CNN) for fake news content [2,3] or fake profile [4] detection, as well as Graph Convolution Networks (GCNs) for the analysis of fake news diffusion paths [5].

The advent of Large Language Models (LLMs) and generative Artificial Intelligence (AI) has added a new dimension to this challenge. These technologies possess the dual capability to both generate and detect fake news, creating a paradox in their application. On one hand, LLMs can be harnessed to create convincingly realistic but false content, making the identification of such fake news increasingly difficult [6]. On the other hand,

these same models offer advanced methods for detecting and combating the spread of fake news [7], presenting a potential solution to the problem they help create.

The importance of detecting fake news cannot be overstated. Reliable detection mechanisms are crucial for maintaining the integrity of information ecosystems, ensuring that the public can trust the information they consume. Effective detection not only helps in identifying false information but also in curbing its spread, thereby mitigating its impact.

This survey paper aims to explore the multifaceted role of LLMs and generative AI in the realm of fake news. It delves into the background of fake news, fake profiles, and the mechanisms of fake news dissemination and astroturfing. Furthermore, it examines the dual role of *LLMs as both generators and detectors of fake news*, highlighting the importance of detection technologies. The paper also sets forth the objectives of this survey, aiming to provide a comprehensive overview of the current landscape and emerging trends in this field.

The structure of this survey is organized as follows: First, in Section 2 we provide a detailed background on fake news, fake profiles, and the state-of-the-art approaches used before LLMs for their detection. Section 3 describes the methodology used in this study for searching and selecting the most relevant articles in the literature. Section 4 provides more details on the different ways that LLMs can be used to generate and detect fake content and profiles in social networks. Section 5 presents the different datasets, metrics, and techniques used to assess the performance of LLMs in fake news detection, and summarizes the main performance results in this task. Section 6 presents case studies and practical applications of the use of LLM in the fake news domain. Section 7 discusses the main challenges and limitations and Section 8 concludes the paper by providing references to future directions in the field.

2. Overview of Fake News and Fake Profiles

2.1. Disinformation/Fake News, Spreading

Fake news and disinformation have become pervasive issues in the digital age, exacerbated by the rise of social media and advanced AI technologies. Understanding the origins, methods, and impacts of fake news is crucial for developing effective detection and mitigation strategies. In the following we explore the evolution of fake news creation and detection, highlighting the role of both human and AI-driven approaches.

Fake news is defined as false news deliberately spread to deceive people, as stated by the Oregon Library [8]. In the pre-LLM era, the creation of fake news was purely the result of human manipulation [9]. Prior to the appearance of LLMs, research in AI-generated misinformation detection focused mainly on Smaller Language Models (SLMs) [10], such as BERT [11], GPT-2 [12], and T5 [13].

Disinformation, which implies an intent to mislead, refers to false information disseminated tactically to bias and manipulate facts, often associated with the term "propaganda". The prefix dis- indicates a reversal or negative instance of information [14].

With the rise of LLMs, fake news can now be both automatically generated and detected [9]. The complexity of disinformation detection has dramatically escalated with the advent of LLMs, which possess billion-scale parameters [10]. In recent years, the popularization of the internet and social networks has facilitated the rapid spread of fake news, causing severe personal, social, and economic damage [15]. Fake news poses significant risks in various areas [6], like public health and healthcare, politics, and e-commerce. The threat of the automatic generation of fake news through large language models (LLMs) is considered one of the most significant risks to the further development of LLMs [16]. Therefore, it is urgent to update detection systems with knowledge of LLM-generated fake news [17].

Fake profiles in Online Social Networks (OSNs) are another significant concern. These profiles are used in Advanced Persistent Threat cases to spread malware or links to it. They are also used in other malicious activities like junk emails, spam, or artificially increasing the number of users in some applications to promote them [18]. Additionally, fake accounts are created to boost the number of fans for celebrities or politicians on social media.

The widespread use of social networking has led to various problems, including the exposure of users to incorrect information through fake accounts, which results in the spread of malicious content [19]. Table 1 presents a comparison of disinformation and/or fake news detection studies.

Table 1. Summary of key studies on Fake News Detection

Ref.	Models Used	Dataset/Data Type	Main Contribution	Best Performance Value	Limitation
[10]	GPT-3.5, GPT-4	LLM-generated disinformation datasets	Explores the efficacy of LLMs in detecting LLM-generated disinformation	Misclassification rate: GPT-4 (all scale)—4.7%	Existing detection techniques struggle with LLM-generated disinformation; relies heavily on advanced prompts
[9]	ChatGPT, OPT-IML-Max, Vicuna, GPT-4, etc.	News articles generated by LLMs	Evaluates disinformation generation capabilities of LLMs and effectiveness of detection models	ChatGPT: 0.97 AUC, 0.82 Macro F1-score	High variability in model performance; detection models may not generalize well across different LLMs
[6]	BERT, RoBERTa, Grover, DualEmo, Llama2-7b, etc.	Fake news detection	Proposes VLPrompt for generating fake news and evaluates different prompting strategies	Llama2-7b + LoRA: ACC 0.971, F1 0.955, PRC 0.986, RCL 0.927	Limited exploration of other prompting strategies; high computational cost for fine-tuning LLMs
[20]	RoBERTa, BERT, ELECTRA, ALBERT, DeBERTa	GossipCop++ and PolitiFact++ datasets	Examines the performance of detectors on different types of news (real, fake, machine-generated)	Large RoBERTa: 99.51% accuracy with 100% machine-generated news	Bias against machine-generated texts; requires balancing dataset composition for optimal training
[14]	Transformer models, classical ML models	Text, audio, video datasets	Proposes a taxonomy for fake news detection and evaluates various feature representation and classification methods	Context-dependent models with transformers perform best	Optimal feature extraction techniques vary with dataset characteristics; limited exploration of audio and video data
[8]	FactAgent utilizing pre-trained LLMs	News claims	Introduces FactAgent for structured workflow in fake news detection	N/A	Dependent on structured workflow and domain-specific tools; does not involve model training
[17]	ChatGPT, various LLMs	Human-written and LLM-generated misinformation	Investigates the detection difficulty of LLM-generated misinformation compared to human-written misinformation	LLM-generated misinformation harder to detect	Detection difficulty varies with misinformation type; requires extensive empirical validation across more models

2.2. Detection Methods

Internet and social media made access to news information much easier and more comfortable. On social media platforms, every user from around the world can publish and disseminate any kind of statement, or set of statements, to spread fake news to achieve different goals, which may be fraudulent or illegal [3]. Aside from individuals spreading

fake news, some sources are considered to be authentic and are popular sources for informational services, such as Wikipedia, which are also prone to false information or fake news [21]. Mass media may also manipulate the information in different ways to achieve their goals. As the sources of spreading fake news are too many, the detection of fake news is essential. One important detection mechanism is the fact-checking sites, but unfortunately, they are insufficient to combat the amount of disinformation disseminated [22]. Fake news is increasingly sophisticated, adopting complex characteristics that differ between different types of social networks. Therefore, it is essential to develop detection mechanisms using artificial intelligence or machine learning algorithms. The paragraphs that follow present the two main fake news detection and mitigation types: (i) Rule-based Approaches and (ii) Human Fact-Checking and Moderation and the key solutions of each type.

2.2.1. Rule-Based Approaches

Rule-based approaches for fake news detection generally fall into two main methods: *Keyword Analysis* and *Heuristic Rules*.

The *Keyword Analysis* method explores the Positive and Unlabeled Learning (PUL) combined with a network-based Label Propagation (PU-LP) algorithm for detecting fake news [23]. This combination reduces labeling costs while improving the classification task by exploiting relationships between news and terms using a few labeled fake news examples. The PU-LP algorithm is integrated with a Graph Attention Neural Event Embedding (GNEE) to further refine the relevance of terms used in fake news detection. Experiments show that such approaches, especially when using specialized keyword extraction tools, achieve high accuracy with minimal labeled data. The potential incorporation of attention mechanisms is also expected to improve effectiveness in fake news detection.

Other studies in this group [3,24] introduce hybrid deep learning models that combine Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) for fake news classification. The models leverage CNN's ability to extract local features and the Long Short-Term Memory (LSTM) network's capability to learn long-term dependencies. By processing input data to capture both local and sequential features, the models classify text as fake or not. Validation on popular fake news datasets (ISO and FA-KES) shows that hybrid models significantly outperformed other non-hybrid baseline methods.

Furthermore, the Interpretable Knowledge-Aware (IKA) model uses historical evidence in the form of graphs for fake news detection [25]. It constructs positive and negative evidence graphs by collecting signals from historical news. IKA compares the given piece to be detected with these evidence graphs to generate a matching vector, enhancing prediction accuracy. Semantic, stylistic, and emotional elements are extracted from news articles and comments and combined with graph-based data for prediction. The model was tested on English and Chinese datasets, achieving high accuracy. IKA is versatile and can be integrated with other existing models to improve their interpretability.

The *Heuristic Rules* approach utilizes artificial intelligence algorithms like the Naive Bayes classifier to detect fake news on social media [26]. Each user post represents a news item, with each word processed independently to calculate the probability of the post being fake. The classification process evaluates the likelihood that the post is fraudulent based on the presence of certain words. To ensure numerical stability, the system uses default probabilities for unknown words and logarithmic computations. This method achieves a classification accuracy of approximately 74%, which is acceptable for the simplicity of the approach.

More advanced deep learning techniques exploit the power of Graph Convolution Neural Networks (GCNs) to analyze the text of user posts and/or user profiles and detect fake news or fake profiles [5]. They also take advantage of the news diffusion paths to further detect hidden attempts to disseminate fake content, by engaging fake and legitimate user accounts [27].

2.2.2. Human Fact-Checking and Moderation

Human fact-checking and moderation approaches play a crucial role in the fight against fake news by leveraging the expertise of individuals and organizations to verify information. These approaches can be broadly categorized into *Manual Verification Processes* and *Fact-Checking Organizations* methods.

Manual Verification Processes involve human experts who assess the credibility of information through thorough investigation and cross-referencing with reliable sources. One of the seminal works on the automatic detection of fake news was conducted by Vlachos and Riedel [28]. The authors defined the task of fact-checking, collected a dataset from two popular fact-checking websites, and used k-Nearest Neighbors classifiers to handle fact-checking as a classification task.

Fact-Checking Organizations are dedicated entities that specialize in verifying the validity of information across various domains. Several well-known organizations include:

- PolitiFact [29]: A well-known non-profit fact-checking website in the United States [30]. It mainly rates the accuracy of claims made by US political news and politicians' statements using classifications such as true, mostly true, half true, mostly false, false, and "pants on fire" (i.e., clearly fake news) [31].
- Snopes.com [32]: One of the first online fact-checking sites, launched in 1994 [31]. The site addresses political, social, and other current issues, primarily confirming news stories spread on social media.
- Suggest (also known as GossipCop) [33]: This site investigates fake stories about US celebrities and entertainment published in magazines and web portals. Each story is rated on an authenticity scale from 0 to 10, where 0 denotes a completely false rumor and 10 indicates a report that is entirely true.
- FactCheck.org [34]: A non-profit site that describes itself as a "consumer advocate for voters" aiming to reduce the level of deception and confusion in US politics [31]. It primarily assesses assertions made by politicians in speeches, debates, and TV commercials.
- FullFact [35]: An online fact-checking site supported by a British charity that verifies and corrects news and assertions circulating on UK social media.

2.3. Fake Profile Detection

Online social networks such as Facebook, X (former Twitter), RenRen, LinkedIn, and Google+ have become increasingly popular over the last few years [36]. OSNs dramatically impact people's lives [37], as they are used to interact with others, exchange information, organize events, engage in e-commerce, and even run their own e-businesses. However, one of the ways people misuse social media is by creating fake profiles for various unethical purposes, such as targeting specific users, spreading misinformation, or engaging in other cybercrimes [38]. Therefore, the identification and elimination of fake accounts on OSNs are essential for maintaining the platforms' integrity and security. In the following, we present the three main methods in this task: Profile picture verification, Network analysis, and Behavior and interaction patterns.

2.3.1. Profile Attribute Analysis

Profile attribute analysis is crucial for distinguishing between genuine and fake social media accounts. This analysis employs two main approaches: *Profile Picture Verification* and *Consistency of Profile Information*. Each method involves various ML techniques and algorithms to ensure the authenticity of profiles, thereby enhancing the security and reliability of social networking platforms.

The *Profile Picture Verification* approach leverages several methods to verify the authenticity of profile pictures. One such method is the Photo-Response Non-Uniformity (PRNU) technique, which can be successfully used as a watermarking approach for authorship attribution and verification of pictures on social networks. According to [39], PRNU employs a unique noise pattern for each camera that withstands the compression algo-

rithms applied by various social platforms, unlike traditional watermarking techniques. This method enables accurate detection of fake accounts without altering the original image's appearance.

Additionally, a new algorithm named SVM-NN is proposed by [36] to provide efficient detection for fake Twitter accounts and bots. This machine learning classification algorithm utilizes fewer features than traditional support vector machine (SVM) or neural network (NN) algorithms. The SVM-NN model involves training an NN model using SVM decision values, effectively distinguishing between Sybil and real accounts.

Moreover, a hybrid model described in [40] uses machine learning and skin detection algorithms to identify fake accounts. This model integrates five supervised machine learning algorithms—K-nearest Neighbor, Support Vector Machine, Naïve Bayes, Decision Tree, and Random Forest—to analyze user account features. It achieves up to 80% accuracy in detecting fake accounts by calculating the percentage of skin exposure in profile images, finding that fake accounts often exhibit higher skin percentages.

The *Consistency of Profile Information* approach focuses on the textual and metadata consistency of profile information. One significant method is the detection of fake LinkedIn profiles immediately after creation, as described by [41]. Utilizing textual data provided during the registration process, the system employs section and subsection tags (SSTE) to detect fake accounts before they establish connections with real users.

Furthermore, a data mining approach for fake profile identification, as detailed by [37], includes four well-known techniques: Neural Network (NN), Support Vector Machine (SVM), Principal Component Analysis (PCA), and Weighted Average (WA). PCA is used for dimensionality reduction, selecting the most explanatory profile features. NN, SVM, and WA predict the authenticity of profiles, with SVM using a Polynomial Kernel achieving the highest accuracy at 87.34%.

Finally, a proposed framework combines the Random Forest Classifier model with pipelining to identify fake profiles across social communities, as described in [42]. The classifier, optimized through grid search, evaluates features like profile picture, bio length, external URL, privacy status, number of posts, followers, and follows. The pipelining process enhances model accuracy by automating the workflow and evaluating the model's performance using a confusion matrix for precise and reliable results.

2.3.2. Network Analysis

In network analysis, various methods are utilized to detect and analyze fake accounts on social media platforms. These methods focus on examining relationships and communities within the network, leveraging graph-based algorithms and community detection techniques to ensure the authenticity of user profiles.

Relationship Graphs is a fundamental aspect of detecting fake users. An enhanced graph-based semi-supervised learning algorithm (EGSLA) has been introduced to detect fake users on Twitter by examining user activity over a prolonged period [43]. EGSLA employs dynamic feature changes, which are challenging for spammers to manipulate over time since spammers typically mimic real users temporarily due to the associated costs. By representing user activity through a graph model, EGSLA classifies, predicts, and makes recommendations using scalable streaming for real-time data processing. The algorithm's performance, tested against existing game theory, k-nearest neighbor (KNN), support vector machine (SVM), and decision tree techniques, achieves a 90.3% accuracy in identifying fake users, outperforming other methods and underscoring its effectiveness.

Another approach proposes a hybrid method for detecting malicious profiles on Twitter by utilizing machine learning techniques. Initially, a Petri net structure analyzes user profiles and features [44], which are subsequently used to train various classifiers to distinguish between malicious and genuine users. The framework processes baseline data through different classifiers in machine learning environments, creating separate datasets for authentic and fraudulent accounts. By applying reachability checks, the system identifies a harmful activity, involving interaction anomaly detection and addressing low

accuracy issues through a trackback mechanism. A comparative analysis demonstrates the framework's efficiency, showing improved performance over state-of-the-art techniques.

Furthermore, a real-time Sybil detector for the Renren online social network has been developed, operating based on a graph where nodes represent user profiles and edges denote connections between profiles through interactions and relationships [45], taking into account the interaction anomalies that may exist. The detector monitors account features such as friend-request frequency, outgoing request acceptance rates, and clustering coefficients to identify suspicious behavior. By analyzing the graph structure, particularly patterns of link creation and clustering coefficients, it distinguishes between genuine and Sybil accounts. If an account exceeds predefined limits, it is flagged as a potential Sybil by a threshold-based system. An adaptive feedback scheme dynamically tunes the threshold parameters, leading to the identification and banning of over 100,000 Sybil accounts.

Community Detection is another critical method for identifying fake accounts. The Communal Influence Propagation Framework (CIPF) identifies fake accounts by analyzing essential features from individual user profiles, linguistic characteristics, and group profiles (network features) [46]. The framework scrutinizes these features to generate a feature vector of the user space, utilizing Social Media Mining (SMM) theories such as Influence, Homophily, and Balance to enrich the malicious user space as an influential index. The Jaccard coefficient evaluates similarity indices, detects Sockpuppet nodes, and generates a negative propagation belonging matrix. CIPF integrates two influence-based verification models—the Sockpuppet Detection Phase (IB-SPD) and the Convolutional Neural Network-based Crowdturfing Community (IB-CFC)—to robustly identify and verify fake accounts.

Additionally, a majority voting approach leverages machine learning algorithms to identify and categorize counterfeit social media profiles [47]. This approach employs various algorithms, including Decision Trees, XG-Boost, Random Forest, Extra Trees, Logistic Regression, Ada-Boost, and K-Nearest Neighbors, each capturing distinct aspects of user behavior and profile attributes. By combining several techniques, a group of classifiers is formed, and a majority vote process determines the legitimacy of a social media profile. This methodology effectively distinguishes authentic user profiles from fake ones by utilizing the decision-making capabilities of multiple classifiers.

Lastly, the Grey Wolf Optimization Algorithm (GWO), inspired by the hunting behavior of grey wolves, finds optimal solutions for complex problems by balancing exploration and exploitation [48]. In this context, each online social network (OSN) user is classified as either "Real" or "Fake". Compared to conventional machine learning algorithms such as K-Nearest Neighbor, Support Vector Model, Decision Tree, and Random Forest, GWO outperforms them in a Facebook dataset, achieving the highest accuracy (0.98), precision (0.96), recall (0.97), and F1-score (0.96). This demonstrates its efficacy in optimizing classification tasks and identifying fake profiles.

2.3.3. Behavior and Interaction Patterns

In the realm of behavior and interaction patterns, various methods are utilized to detect and analyze fake accounts on social media platforms. These methods focus on examining the frequency of posts, the content of tweets, and the interaction patterns of users, employing machine learning algorithms to identify anomalies and ensure the authenticity of user profiles.

Posting Frequency/Tweets are critical indicators in detecting fake accounts. A study presents a classification method for detecting fake accounts on Twitter using the Naïve Bayes algorithm [19]. The data include features provided by the Twitter API, such as description, profile image, number of followers, users, and public lists in which the user participates. Additionally, they consider the number of hashtags, mentions, and URL links used in the last 20 tweets, as well as the number of tweets liked by the user. The dataset was preprocessed using a supervised discretization technique named Entropy Minimization Discretization (EMD) on numerical features. In the Naïve Bayes learning

algorithm, preprocessing the data with EMD without any feature selection increases the success rate, as Naïve Bayes performs better with discrete values than continuous values. The success rate of the Naïve Bayes algorithm is 90.9% after discretization.

In another study, Facebook user feeds were collected using the Facebook Graph API, which was fine-grained based on privacy settings [49]. The most commonly used supervised machine learning classification techniques were applied to a basic dataset comprising the user's node, their friends in their social neighborhood, and a set of manually identified spam accounts. The performance of the classifiers was evaluated to determine their ability to detect fake accounts on Facebook. Results show that the best-performing classifiers achieved a high classification accuracy of 79%. Furthermore, the study investigated the feature set to determine the attributes that most successfully detect fake profiles on Facebook. It was found that user activities related to likes, comments, and, to some extent, shares on Facebook contribute the most to detecting fake accounts. Therefore, the study represents a significant step towards a profile-feature-based detection of fake accounts on Facebook.

3. Methodology

To determine the related scope of the current work, we utilized the PRISMA 2020 guidelines [50]. These guidelines outline the systematic approach adopted for reviewing literature on the usage of LLMs for generating fake news, and fake profiles, and also the methods of detection based on LLMs.

Search Strategy: The initial search was conducted across multiple academic sources, including Google Scholar and IEEE Xplore, but mainly focused on Scopus due to the better quality of related results and improved control during the filtering process. In Scopus, we used a set of search parameters that look for specific keywords in the titles, abstracts and keywords, and limit the search range to works published from 2018 onwards. We targeted published works in journals, conferences, and review papers. The query used for Scopus is presented below.

```
TITLE-ABS-KEY(("fake news" OR "misinformation" or "disinformation") AND ("llm
    " OR "large language models"*)) AND PUBYEAR > 2018 AND PUBYEAR < 2025 AND
     (LIMIT-TO( DOCTYPE , "re") OR LIMIT-TO (DOCTYPE , "ch") OR LIMIT-TO (
    DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "cp") ) AND (LIMIT-TO (LANGUAGE ,
    "English"))
```

Search Criteria: To gather relevant works, we extended our search to include Google Scholar and IEEE Xplore. In addition to the initial keywords, we used composite phrases such as "LLM in fake news generation", "fake profile detection using LLM", and "security challenges in detecting fake news with LLM". We conducted several queries and manually validated the results to find more relevant papers, including pre-prints. By screening abstracts, assessing the credibility of sources, and prioritizing peer-reviewed articles, we selected works that address the various aspects of using LLMs for generating and detecting fake news and profiles.

We also employed a snowballing technique, which involves examining the references of initially identified papers to discover additional relevant works. This method allowed us to uncover papers that might not have surfaced during our initial search. By exploring citations within these papers, we could track the development and interconnections of research topics over time. We also utilized forward snowballing, looking at newer papers that cited our core set of articles. Iterative refinement of our search strategies further helped us in this process. The methodology we followed in selecting papers for inclusion in the survey is depicted in Figure 1. The results of our literature review are detailed in the following sections.

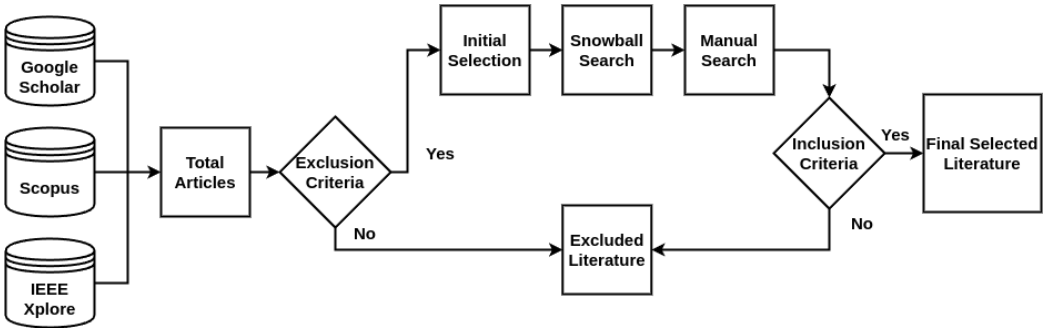

Figure 1. Flowchart illustrating how papers were systematically selected for the survey.

Inclusion and Exclusion Criteria: The search was restricted to English-language articles. We limited the search to works published in the years after 2018 and further filtered the papers using additional criteria. The inclusion criteria were as follows: peer-reviewed articles focusing on using large language models (LLMs) to generate fake news, detect fake profiles, and address security challenges associated with these activities. We also considered papers that discuss various techniques for improving the effectiveness and accuracy of LLMs in these contexts. Exclusion criteria included non-peer-reviewed articles and publications not specific to LLMs or their application in fake news and profile detection. In total, 117 publications met the inclusion criteria. The distribution of these publications over the years is presented in Figure 2, while Figure 3 shows the types of the accepted publications.

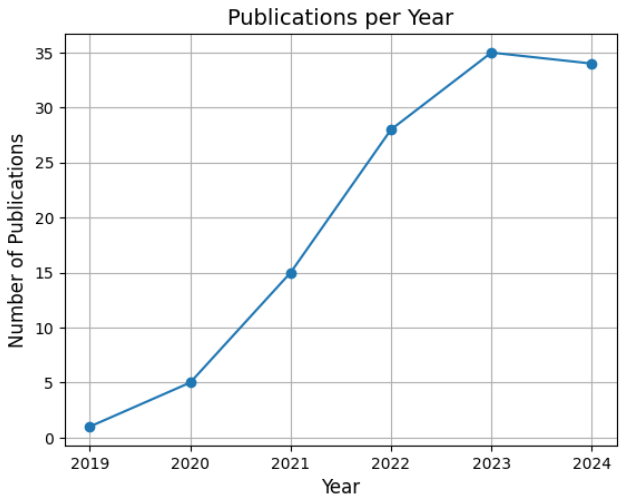

Figure 2. Retrieved publications per year.

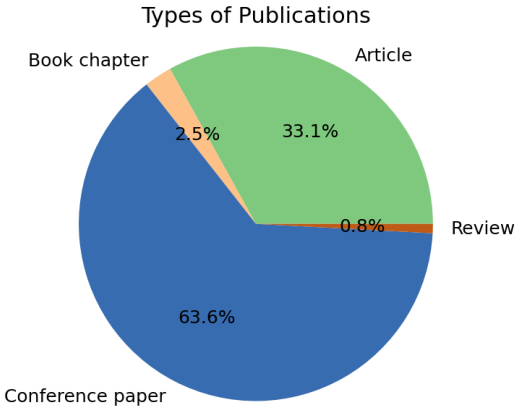

Figure 3. Number of publications per type.

Quality Assessment: The quality of the selected papers was meticulously evaluated based on several criteria: their relevance to the research topics, methodological rigor, the impact factor of the journals in which they were published, and their citation counts. This comprehensive evaluation ensured that only the most credible and significant studies were included in our survey, thereby enhancing the reliability and depth of our findings.

Data Extraction and Synthesis: Data extraction and synthesis focused on identifying categories and methods related to the use of LLMs for generating fake news, creating fake profiles, and detecting these activities. Figure 4 depicts the most popular topics mentioned in the selected articles. We specifically considered survey works that illustrate different approaches to using LLMs, studies discussing aspects such as detection accuracy, robustness of fake news generation, and user identification techniques. Given the recent broad access to LLM technologies, the number of surveys and studies available was limited. This limitation was reflected in the literature, which mostly consists of recent publications. The limited literature on LLMs and their application in fake news and profile detection underscores the emerging nature of this field and the need for further research and exploration.

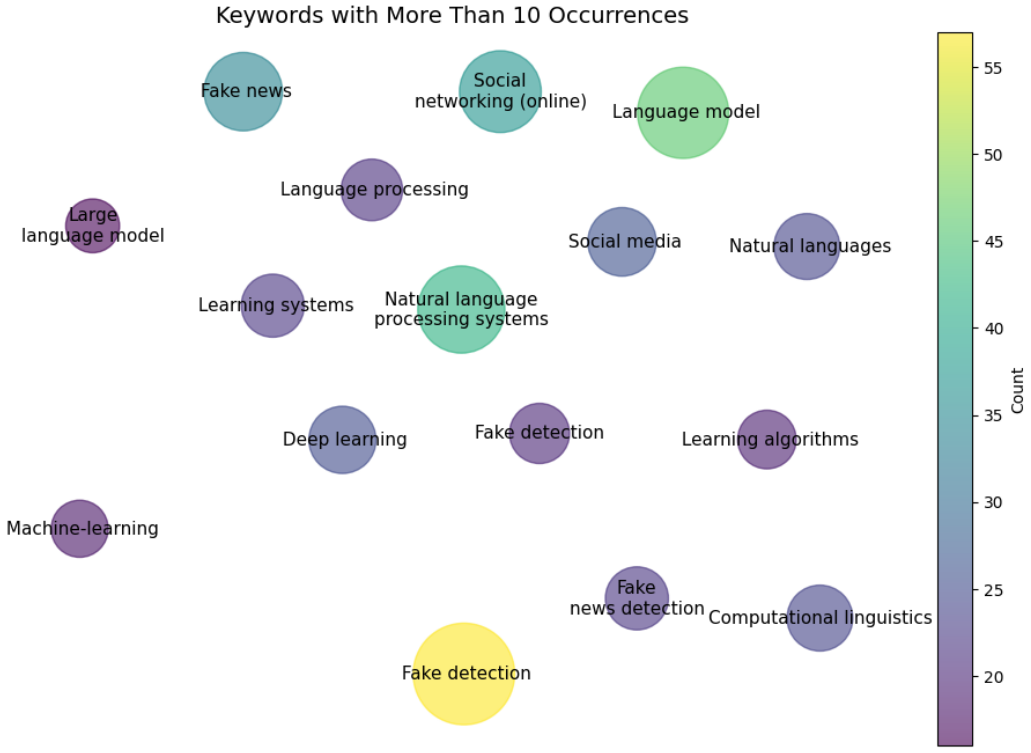

Figure 4. The most popular terms in the titles of the retrieved articles. Higher term count values correspond to a bigger size of the corresponding bubble and a lighter color.

Results of the Literature Search: The findings are organized in alignment with the document's structure, providing a detailed analysis of the use of large language models (LLMs) for generating fake news, creating fake profiles, and detecting such activities. This includes examining the existing literature that highlights the importance and various dimensions of LLMs in the context of fake news and profile detection. Given the limited literature due to the recent broad access to LLM technologies, our survey covers the most relevant and recent studies in this emerging field. We analyze the different methodologies employed, the effectiveness of various detection techniques, and the security and privacy considerations associated with using LLMs for these purposes.

4. Methods That Use Large Language Models

LLMs are sophisticated AI models designed to understand and generate human language. They have evolved significantly, enabling a wide range of applications, including the potential for malicious uses such as generating fake news and fake profiles, as well as spreading misinformation. However, these unique abilities also provide powerful tools for detecting such misuse. In Section 4.1, we explore some of the most prominent LLMs and their contributions to the field.

The advancements in LLMs have led to their application in both the generation and detection of fake news and fake profiles. The following sections will delve into these specific applications, discussing how LLMs are used to generate fake news and profiles, and how they can be leveraged to detect such deceptive content effectively. These advancements will be explored in the Sections 4.2–4.4.

4.1. Introduction to Large Language Models (LLMs)

During the last few years, an abundance of pre-trained models has been developed and used in various linguistic tasks, boosting the performance of traditional methods, and using only fine-tuning in most cases. In this section, we present the details of these models, whereas Table 2 provides a summary of the most frequently used models.

The evolution of large language models began with BERT (Bidirectional Encoder Representations from Transformers), introduced by Google researchers in 2018. BERT is a language representation model that learns language in a bidirectional manner. It has two variants: BERT-BASE (12 layers, 110 M parameters) and BERT-LARGE (24 layers, 340M parameters). BERT was trained on the BooksCorpus (800 million words) and English Wikipedia (2.5 billion words), excluding lists, tables, and headers [11]. Building on BERT, M-BERT (Multilingual BERT) was released in 2019 as a single language model pre-trained on Wikipedia pages of 104 languages. It has a 12-layer transformer architecture and comes in several variants: BERT-Base Multilingual Cased (104 languages, 110 M parameters), BERT-Base Multilingual Uncased (102 languages, 110M parameters), and BERT-Base Chinese (110 M parameters) [51].

Another significant model is T5 (Text-to-Text Transfer Transformer), released in 2019, which is an encoder-decoder model that converts each task into a text-to-text format. It was trained on the Colossal Clean Crawled Corpus (C4) dataset, a 750 GB dataset created from Common Crawl's web-extracted text. The model architecture is similar to the original Transformer, with both encoder and decoder consisting of 12 blocks, and it has 220 million parameters [13]. Following T5, RoBERTa is an enhanced version of BERT with more data, dynamic masking, and byte-pair encoding. RoBERTa improves performance by training longer, using larger batches, removing the next sentence prediction objective, and dynamically changing the masking pattern. RoBERTa was trained on a large dataset called CC-NEWS [52].

In the realm of models that are fine-tuned for specific tasks, FN-BERT stands out as a BERT-based model fine-tuned on a fake news classification dataset in 2023. The model is based on distilBERT and was trained using a specific fake news dataset [53]. Meanwhile, Grover, developed by the Allen Institute for AI, is a GAN-based model designed to generate and detect fake news. Grover has three sizes: Grover-Base (117 M parameters), Grover-Large (345 M parameters), and Grover-Mega (1.5 B parameters). It was trained on the RealNews dataset, created from Common Crawl news articles [54].

Recent advancements include Llama2, a collection of pre-trained and fine-tuned LLM models from Meta with varying parameter sizes. It includes Llama 2 Chat, optimized for dialogue, and Code Llama, specialized for code generation. Llama2 was trained on 2 trillion tokens, with increased context length and grouped-query attention [55]. Additionally, Meta has developed the Llama3 models, which show significant improvements in performance and scalability compared to their predecessors.

Microsoft's Phi3 [56] also presents significant advancements in performance and scalability. It emphasizes improvements in computational efficiency and model adaptability,

making it a strong contender in the evolving landscape of large language models. Phi-3 comes in different versions ranging from 3.8 B (Phi-3 mini) to 14 B (Phi-3 medium) parameters, which offer two context lengths, 128 K and 4 K, and outperform bigger models. They have been trained on 3.3 million tokens.

OpenAI's GPT-3 demonstrated remarkable abilities in translation, question answering, and content generation, comparable to human-written texts. Furthermore, ChatGPT-3.5, also developed by OpenAI, is widely used for its conversational abilities. It was trained using Reinforcement Learning from Human Feedback (RLHF) and fine-tuned from a model in the GPT-3.5 series. ChatGPT can answer follow-up questions, admit mistakes, and reject inappropriate requests [57]. Continuing this progress, OpenAI's GPT-4, released in March 2023 [58], opened new horizons by enabling the understanding and answering of questions from both text and image inputs, fostering a more detailed and comprehensive understanding of information. This progress in the GPT family has raised expectations and broadened the range of their applications.

Alibaba's Qwen2 [59], another advanced model, focuses on enhancing accuracy and efficiency in natural language processing. Mistral AI's Mistral models [60] and Google's Gemini models [61] also contribute to expanding the capabilities of open standards by providing advanced solutions for complex learning problems.

Table 2. Overview of Key Large Language Models (LLMs).

Model	Tokens	Parameters	Training Corpus Size	Architecture	Reference
BERT	3.3 B	BERT-BASE: 110 M, BERT-LARGE: 340 M	BooksCorpus (800 M words), English Wikipedia (2.5 B words)	Transformer (12/24 layers)	[11]
M-BERT	300 M	110 M	Wikipedia pages of 104 languages	Transformer (12 layers)	[51]
T5	500 M	220 M	C4 (750 GB)	Encoder-Decoder (12 blocks)	[13]
RoBERTa	500 M	355 M	CC-NEWS, Web texts (160 GB)	Enhanced BERT	[52]
FN-BERT	300 M	66 M	Fake news dataset	Fine-tuned DistilBERT	[53]
Grover	400 M	Grover-Base: 117 M, Grover-Large: 345 M, Grover-Mega: 1.5 B	RealNews dataset (120 GB)	GAN-based Transformer	[54]
Llama2	2T	Llama2-7B, Llama2-13B, Llama2-70B	Various datasets	Transformer	[55]
Llama3	15 T	8 B, 70 B	Massive dataset (incl. Common Crawl, code, data, books)	Improved Transformer	-
Phi3	3.3 M	Small 3.8 B, Medium 14 B	Large-scale dataset	Improved Transformer	[56]
GPT-3	500 B	175 B	Diverse datasets (web, books, Wikipedia)	Multimodal Transformer	[58]
GPT-4	10 T	1.7 T	Diverse datasets (web, books, Wikipedia)	Multimodal Transformer	[58]
Qwen2	500 B	200 B	Massive dataset	Transformer	[59]
Mistral	500 B	175 B	Large-scale dataset	Transformer	[60]
Gemini	1 T	1.4 T	Massive dataset	Advanced Transformer	[61]

4.2. LLM-Based Fake News and Fake Profiles Generation

The generation of fake news and fake profiles has evolved significantly with the advent of LLMs. Traditionally, the automatic creation of fictitious news pieces relied on simplistic methods such as word jumbles and random replacements within actual news articles [6,54,62]. These rudimentary techniques often resulted in content that was easily recognizable as fake due to its lack of coherence.

With the introduction of LLMs, there has been a substantial shift in the research focus towards creating more coherent fake news. Initial efforts involved using prompting techniques to generate fake news, which aimed to produce more believable narratives [63,64]. However, these generated stories were still detectable by other language models because they often lacked sufficient details or consistency.

To address these shortcomings, researchers introduced methodologies that combined real news, factual information, and intentionally false information provided by humans. For example, Su [20] tasked LLMs with fabricated articles from human-collected summaries of fake events. Similarly, Wu and Hooi refined the generation of fake news articles using LLMs to enhance their believability [64].

Other innovative approaches include Jiang's method, which uses fake events combined with real news articles to generate fake news [10], and Pan's approach, which employs a question-answer dataset based on real news to manipulate answers and create fake news content [6,65]. These strategies, particularly those involving manually crafted fake news, help prevent the automated mass production of fake news articles. Nonetheless, techniques that rely on fabricated summaries often produce content lacking in detail, and any alterations to specific events or elements can lead to issues with contextual coherence.

4.3. LLM-Based Fake News Detection

Detecting fake news has become increasingly sophisticated with the integration of large language models (LLMs). Traditional fake news detection methods often rely on auxiliary data in addition to the article's text. For instance, mainstream detection programs use metadata such as authors, publishers, and publication dates to determine the authenticity of an article. A notable example is Grover, which leverages these metadata for validation [54]. Similarly, DeClare compares statements to data collected from web-searched articles to verify their credibility [66]. Another approach involves extracting emotional and semantic characteristics from texts, as demonstrated by Zhang and colleagues in 2021.

Moreover, platforms like Defend examine the responses garnered by publications on social media to aid in fake news detection [67]. However, these methods have limitations in practical applications. Auxiliary data are not always accessible, and relying on online fact-checking services like PolitiFact and Snopes is labor-intensive, making it difficult to keep up with the rapid production of misinformation [68].

Recent studies have shown that fine-tuned pre-trained language models (PLMs) such as BERT, and LLMs like GPT-3.5, can achieve commendable results in fake news detection without the need for extensive auxiliary data [6,20,63,64]. The primary approaches for LLM-based fake news detection include text classification, fact-checking and verification, and contextual analysis, which will be analyzed in the following subsections.

4.3.1. Fake Text Classification Using LLMs

In recent years, the fine-tuning of pre-trained models on fake news datasets has emerged as a pivotal approach in enhancing the accuracy of fake news detection. This method leverages the robust capabilities of pre-trained models and adapts them specifically for identifying misinformation, leading to significant improvements in performance metrics.

Boissonneault et al. fine-tuned the ChatGPT and Google Gemini models using the LIAR benchmark dataset [69]. The training set from the LIAR dataset was employed, utilizing high-performance computing resources, including GPUs, to meet the computational

demands and ensure efficient processing. During the fine-tuning process, hyperparameters such as learning rate, batch size, and the number of training epochs were systematically adjusted to achieve optimal performance. The test set of the LIAR dataset was used for evaluating the fine-tuned models, comparing their predictions to true labels to calculate performance metrics like accuracy, precision, recall, and F1 score. Google Gemini exhibited slightly better results compared to ChatGPT, as evidenced by higher accuracy, F1 score, and superior performance on the AUC-ROC curve.

Al Esawi et al. explored deep learning techniques for detecting Arabic misinformation by leveraging the contextual features of news article content [70]. This approach employed a combination of BiLSTM and the attention mechanism inherent in the transformer architecture. A pre-trained AraBERT model was used to extract features from Arabic text, generating contextual embeddings. These embeddings were then fed into a BiLSTM layer, and the outputs were processed through an attention layer to create a context vector. This context vector was further refined through dense layers and a dropout layer to prevent overfitting, enabling the model to classify each article as fake or real.

Zhang et al. applied fine-tuning of the BERT model to create a health misinformation detection classifier [71]. This classifier processes input text through a transformer architecture comprising multiple layers of bidirectional self-attention mechanisms. The model was trained, validated, and evaluated using three publicly available datasets: CMU-MisCOV19, CoAID (COVID-19 Health Misinformation Dataset), and FakeHealth. The training data were labeled, and the pre-trained weights were adjusted to optimize performance for the task of distinguishing between real and fake health news.

Liu et al. proposed a novel framework named FakeNewsGPT4, which augments Large Vision-Language Models (LVLMs) with forgery-specific knowledge for manipulation reasoning while inheriting extensive world knowledge as a complement [72]. This framework requires two types of forgery-specific knowledge: semantic correlations between visual and textual data and artifact traces indicating manipulations. It employs a multi-level cross-modal reasoning module and a dual-branch fine-grained verification module to extract this knowledge. The generated knowledge is then converted into embeddings compatible with LVLMs using candidate answer heuristics and soft prompts to enhance input informativeness. Experiments have demonstrated that FakeNewsGPT4 outperforms previous methods in cross-domain fake news detection tasks.

4.3.2. Fact-Checking and Verification

Automated fact-checking and verification using large language models (LLMs) has shown considerable promise in improving the accuracy and efficiency of identifying misinformation. Various innovative approaches have been developed to leverage the capabilities of LLMs in this domain.

One approach, named Fine-grained Feedback with Reinforcement Retrieval (FFRR), aims to enable more informed and accurate evidence retrieval for LLM-based fact-checking on real-world news claims [73]. This method involves generating intermediate questions from various perspectives of a claim through prompting to retrieve relevant documents. FFRR then collects fine-grained feedback from the LLM on the retrieved documents at both document-level and question-level, using this feedback as rewards to refine the list of retrieved documents and optimize the retrieval policy for intermediate questions. The FFRR approach, evaluated on two public datasets, significantly outperforms state-of-the-art LLM-enabled and non-LLM baselines.

Another method, known as Hierarchical Step-by-Step (HiSS) prompting, directs LLMs to decompose a claim into several subclaims and verify each of them step-by-step through multiple question-answering sessions [68]. This technique addresses two main challenges in news claim verification: omission of necessary details and fact hallucination. The process includes claim decomposition, subclaim verification, and final prediction. Initially, the LLM decomposes the claim into subclaims to ensure every explicit point is addressed. It then verifies each subclaim by generating and answering a series of insightful questions,

consulting external sources like Google Search when necessary. After verifying the sub-claims, the LLM makes a final prediction and classifies the original claim. Experiments on two public misinformation datasets show that HiSS prompting outperforms state-of-the-art fully-supervised approaches and strong few-shot ICL-enabled baselines.

Another innovative approach uses LLMs to produce weak labels for 18 credibility signals through prompting [74]. Initially, an open-ended prompt generates answers to questions regarding the news article text, which are then mapped to predefined classes using a task-agnostic restrictive prompt. For credibility signals, an instruction prompt is used alongside the article text, followed by a credibility signal prompt. This process is repeated for each credibility signal. The zero-shot prompt is used to detect misinformation as a binary classification task. If simple string matching rules fail to map the model's answer to a label, a task-agnostic category mapping prompt is employed. This method, which combines zero-shot LLM credibility signal labeling and weak supervision, outperforms state-of-the-art classifiers on two misinformation datasets without using any ground-truth labels for training.

4.3.3. Contextual Analysis

Understanding the context of news articles is crucial for effective fake news detection. Various innovative approaches have been developed to leverage the capabilities of large language models (LLMs) for contextual analysis in this domain.

In the work by Hu et al., the Adaptive Rationale Guidance (ARG) network is proposed for fake news detection by combining the advantages of small and large language models [75]. The goal of ARG is to enable small language models (SLMs) to select useful rationales as references for final judgments. To achieve this, ARG encodes the inputs using SLMs, which utilize LLM-generated informative rationales on news content from various perspectives, such as textual description, commonsense, and factuality, to improve performance. The small and large LMs collaborate via news-rationale feature interaction, LLM judgment prediction, and rationale usefulness evaluation, allowing for rich interaction between news and rationales. The obtained interactive features are aggregated to make a final judgment. Additionally, a rationale-free version, ARG-D, was developed for cost-sensitive scenarios without querying LLMs, demonstrating the superiority of both ARG and ARG-D in experiments.

Wu et al. introduced SheepDog, a style-agnostic fake news detector robust to news writing styles [76]. This detector achieves adaptability through LLM-empowered news reframing, which customizes each article to match different writing styles using style-oriented reframing prompts. The style-agnostic training enhances the detector's resilience to stylistic variations by maximizing prediction consistency across these diverse reframings, focusing on content rather than styling.

Wan et al. developed DELL, a framework that identifies three key stages in misinformation detection [77]. First, LLMs generate synthetic reactions to new articles to represent diverse perspectives and simulate user-news interaction networks. Second, LLMs generate explanations of news articles for proxy tasks, refining feature embeddings with these explanations. Finally, DELL employs three LLM-based strategies to selectively merge the predictions of task-specific experts, enhancing calibration. Extensive experiments on seven datasets with three LLMs demonstrate that DELL outperforms state-of-the-art baselines by up to 16.8% in macro F1-score.

Table 3 presents a comprehensive comparison of studies that employ LLMs for Fake News Detection.

Table 3. Comparison of Studies on Fake News Detection.

Study	Models Used	Dataset/Data	Main Advantage	Best Performance Value	Limitations
[75]	Fine-tuned BERT, GPT-3.5, ARG, ARG-D	Two real-world datasets	ARG and ARG-D outperform baseline methods	Macro F1: 0.784	LLM's inability to select and integrate rationales properly
[78]	ERNIE, KnowBert, KEPLER, K-ADAPTER	LIAR, COVID-19	Knowledge-enhanced models improve detection on LIAR	KnowBert-W+W on LIAR: 28.95%	Mixed results on COVID-19, computational cost
[79]	BERT, LSTM, Gradient Boosted Tree	NewsFN data	Fine-tuned BERT model performs well with minimal text pre-processing	Accuracy: 97.021%	Limited comparative context
[80]	Logistic regression, Decision tree	GossipCop++, PolitiFact++	Mitigation strategy improves detection accuracy for human and LLM-generated news	F1 score (PolitiFact++): 0.83	Bias in existing detectors
[8]	FactAgent	PolitiFact, GossipCop, Snopes	Transparent, structured workflow without training process	F1 (PolitiFact): 0.88	Initial implementation without extensive real-world validation
[69]	ChatGPT, Google Gemini	LIAR	High performance in real-world applications	AUC-ROC: High (exact value not specified)	Needs ongoing refinement for optimal performance
[81]	Neural networks with explainable framework	Two real-world benchmarks	Provides high-quality justifications	Outperforms SOTA baselines (exact values not specified)	Dependent on the quality of crowd-sourced data
[82]	RoBERTa	Fake News Challenge Stage 1 (FNC-I)	State-of-the-art performance in stance detection	Weighted accuracy: 90.01%	Limited to stance detection
[20]	RoBERTa, BERT, ELECTRA, ALBERT, DeBERTa	GossipCop++, PolitiFact++	Adapts to LLM era, robust performance on mixed news	Accuracy (GossipCop++): 98.22 (ALBERT)	Detector bias against machine-generated texts
[83]	M-DRUM, GPT-4, LLaVA	Human-centric and Fact-related Fake News (HFFN)	Multi-modal detection and reasoning	Outperforms SOTA models	Limited by the benchmark dataset
[84]	Llama model	70,000 instructions dataset	Parameter-efficient fine-tuning	High accuracy (exact value not specified)	Hard to distinguish deepfake images

4.4. LLM-Based Fake Profile Detection

Profile Content Analysis

MedGraph is a multi-head attention-based graph neural network (GNN) approach to detect malicious behaviors, i.e., edges from a temporal reciprocal graph [85] in online dating applications, considering both the user's features and their interactive behaviors. The method is based on four components. The first component is the motif-based GNN component that defines reciprocal user features through a bipartite graph and uses a motif-based random walk algorithm to sample neighbors based on predefined motifs. A temporal user behavior embedding component discovers abnormal user behaviors based on feature representations of historical interaction data. Then, a co-attention component captures the interactive behavior attributes between users, especially focusing on distinguishing the behavior of abnormal users. Finally, there is the prediction layer to predict if an interaction (edge) is malicious.

4.5. Multi-Modal Approaches

Ayoobi et al. [41] proposed an approach for detecting fake and Large Language Model (LLM)-generated profiles in the LinkedIn Online Social Network immediately upon registration, and before establishing connections with other users. The approach focuses on the textual and metadata consistency of profile information provided during the registration process and leverages advanced word embeddings from models like BERT or RoBERTa.

The approach proposed in [86] integrates a profanity detector and gender identification module for fake profile detection, where the profanity detector identifies the level of profanity in the given text -in this case in the comments posted from the user account-, while the gender detector enhances feature accuracy. It uses pre-trained BERT in combination with Logistic Regression to classify profiles as fake or real. The proposed framework considers an optimized set of 12 features in total, which combinate user features, network-based features, and attribute-based features [86].

5. Comparative Analysis of Traditional and Modern Methods

5.1. Performance Metrics

There is a large variety of evaluation metrics used in the task of fake news or fake profile detection. The metrics consider fake news or profile detection as a binary classification task. Table 4 summarizes the most widely used evaluation metrics.

Table 4. Evaluation Metrics for fake news classification [69].

Metric	Description
Accuracy	The proportion of correctly classified statements, providing a general measure of the models' performance.
Precision	The ratio of true positive predictions to the total positive predictions, measuring the models' ability to correctly identify true statements without generating false positives.
Recall	The ratio of true positive predictions to the total actual positives, assessing the models' ability to detect all true statements, capturing their sensitivity to true information.
F1 Score	The harmonic mean of precision and recall, offering a balanced measure of the models' performance, considering both false positives and false negatives.
AUC-ROC	The area under the receiver operating characteristic curve, providing insights into the models' discriminative ability across different classification thresholds.

In the subsections that follow, we briefly present the performance of various methods discussed in the previous sections in the fake news detection task in various datasets. Where available, the links to the datasets are provided, to allow researchers to further evaluate or improve the methods' performance. Although the comparison of different methods across different datasets is not straightforward, the performance metrics give us an idea of the difficulty of the binary classification task.

5.2. Traditional Methods

5.2.1. Fake News Detection: Rule-Based Approaches—Keyword Analysis and Heuristic Rules

KAFN-PULP [23] is a method that operates with few labeled samples (news items). It extracts keywords from the text, constructs a network that contains a few fake and real labeled nodes among all other nodes, and then applies the Graph Attention Neural Event Embedding (GNEE) method to classify unlabeled nodes. The method outperformed its competitors on the FakeNewsCorpus dataset

The authors in [24] proposed an attention-based ANN with textual, social and image information sources and applied it to the Twitter and Weibo datasets, achieving an accuracy of 75%. The authors in [25] conducted experiments on both Chinese (using the Weibo21 dataset) and English real-world datasets, with the IKA model. They achieved an accuracy of 90.06% on the English and 93.17% on the Chinese dataset.

The performance of the Naive Bayes classifier in fake news detection [26] on the BuzzFeed News dataset was lower, with a precision of 71% and a recall of 13% due to the skewness of the data in the test dataset.

5.2.2. Fake Profile Detection Using Profile Attribute Analysis

In the task of user profile classification, the SVM-NN method introduced in [36] has achieved a classification accuracy of 98% in detecting fake Twitter accounts and bots, using the MIB dataset.

Moreover, the hybrid model described in [40] achieved an accuracy of 80% in a balanced dataset of 400 Facebook profiles. The accuracy was almost the same when Random Forest, Naive Bayes and Decision Tree classifiers were used, and only the KNN algorithm had a lower accuracy score (at 60%). The authors calculated the percentage of skin exposure in profile images, finding that fake accounts often exhibit higher skin percentages.

In [37], four well-known data mining techniques were applied, Neural network (NN), Support vector machine (SVM), and Weighted average (WA) to the problem of fake profile identification in LinkedIn. The accuracy for NN was 79.75% when all the features were used and 82.83% using selected features. Respectively, SVM with RBF Kernel achieved an accuracy of 79.24% and 81.83% before and after feature selection. SVM using a Polynomial Kernel achieved the highest accuracy at 87.34%. Finally, simple WA achieved an average accuracy of 76.16% when the datasets contain all features, and 74.68% when the datasets contain selected features.

The authors in [42] used two fake account datasets, one from Weibo and one from freelancer.in. Using the Random Forest Classifier model combined with pipelining they identified fake profiles in the two datasets with an F1-score of 93% and 95%, respectively.

5.2.3. Fake Profile Detection Using Network Analysis

The authors in [43] used an enhanced graph-based semi-supervised learning algorithm (EGSLA) and graph-based classification techniques to distinguish fake users from a pool of Twitter users. The performance of their approach was tested on a Twitter dataset comprising 21,473 users and 2,915,147 tweets and they achieved an average F1-measure of 88.95%, outperforming k-nearest neighbor (KNN), support vector machine (SVM), and decision tree (DT) classifiers.

The performance of a fake profile detector framework that builds upon PetriNets [44] is evaluated using a crawled Twitter dataset comprising 6824 Twitter profiles and 59,153,788 tweets. The method used various classifiers at the final classification step like Random forest, Bagging, JRip, J48, PART, Random tree, and Logistic regression. Logistic regression achieved the highest precision score, which is 99.2%, and Random Forest and Bagging followed with 98.9%.

In [47], the majority voting approach and other traditional classification methods were tested once again on the MIB dataset of fake profiles. The AdaBoost classifier outperformed all other methods (i.e., 99.22%), whereas the majority voting approach came second with a performance that compared to that of the Random Forest classifier (i.e., 99.12%).

The gray wolf optimization (GWO) method in [48] was tested on a Facebook profile dataset, which was handcrafted to comprise 1043 real and 201 fake accounts and 14 attributes in total (name, profile picture, likes, groups joined, number of friends, educational status, work, living place, check-ins, etc.). The proposed GWO method achieved an accuracy of 98%, compared to the 97% achieved by DTs, SVMs and KNN, making it a robust choice for identifying fake profiles on social networks. It also achieved an F1-score of 96%,

outperforming conventional machine learning algorithms such as K-Nearest Neighbor, Support Vector Model, Decision Tree, and Random Forest.

5.2.4. Fake Profile Detection Using Behavior and Interaction Patterns

The performance of all seven baseline classifiers (J48, Random Forest, Random Tree, REP Tree, OneR, JRip) was evaluated in [49] on a dataset collected using the Facebook API. The dataset comprised 549 real users who are the direct friends of the authors and 2672 more users, friends of the authors' friends, who are assumed to be real users. The dataset also contained 230 Facebook spammers and black market users, and another 1257 users from black market services who are also assumed as fake users. The methods were evaluated on a subset of the original dataset and the proposed method achieved an accuracy score in the range of 75–80% and an error rate of 20–25%. Only Naive Bayes gave a lower accuracy score (58%) but could detect 92% of the actual fake profiles, which was the best performance among all models.

A work on Twitter accounts [19] collected a dataset of 501 fake and 499 real accounts, each account described with 16 features, including: description length, protected account, number of followers, number of following, number of tweets and retweets, number of favorite tweets, number of public lists participating in, verified account, profile background image, contributor mode enabled, etc. The success rate of the Naïve Bayes algorithm tested on this dataset is 80.6% before and 90.9% after discretization.

5.3. LLM-Based Methods

5.3.1. Fake News Classification Using LLMs

In [69], Boissonneault, et al. fine-tuned the ChatGPT and Google Gemini models using the LIAR benchmark dataset. Google Gemini exhibited slightly better results compared to ChatGPT, as evidenced by higher accuracy, F1-score and superior performance on the AUC-ROC curve.

The authors in [70] used a pre-trained BERT model and BiLSTM to detect fake content on the ArCOV19-Rumors dataset from Twitter and on the AraNews dataset and achieved an accuracy of 93.0% and 88.1%, respectively. They also applied the Att-BiLSTM model, which boosted the accuracy to 96.9% and 90.5%, respectively, in the two classification tasks.

In [71] the BERT model was trained on different pre-processed data of the CMU-MisCOV19 dataset. The first case involved using the raw tweet data with an accuracy of 84.1% and the second using tweets without emojis with an accuracy of 87%. The third case uses tweets with emojis and achieves an increased accuracy of 87.5%. The authors also combined the BERT model with different socio-contextual information and normalization of numerical data in association with cleaned textual tweets, achieving an accuracy above 92%. The results demonstrate that incorporating socio-contextual information, including user and tweet information, enhances the model's performance.

In the case of multi-modal content, the vision-enhanced framework of FakeNews-GPT4 [72] outperformed all the methods it compared against, such as MFND and HAM-MER, showing a 7.7% increase in AUC when tested across different subsets, in the single-domain. In multi-domain the performance of FakeNewsGPT4 remains higher than PandaGPT and HAMMER. The authors employed the DGM benchmark dataset for their experiments.

5.3.2. Fake News Detection via Using LLMs for Fact-Checking and Verification

The FFRR method of [73] was evaluated on two public datasets (the RAWFC and the LIAR-RAW, significantly outperforming other state-of-the-art models. FFRR on the document and question level achieves the highest F1-score (57%) on the RAWFC dataset and the second-best F1-score (33.5%) on the LIAR-RAW dataset. The HiSS method, presented in [68], achieved a lower F1-score (53.9%) on the RAWFC dataset, but achieved a better F1-score (37.5%) on the LIAR dataset.

In [74] different LLMs were used to produce weak labels for 18 credibility signals through prompting, achieving different scores. The highest results on the EuvsDisinfo dataset were achieved using GPT-3.5-Turbo-FULL (99.0% F1-score), whereas the OpenAssistant-30B-FULL model achieved 54.8% F1-score on the FA-KES dataset.

5.3.3. Fake News Detection Using LLMs and Contextual Analysis

The ARG method proposed in [75] achieved a 78.6% accuracy in the Chinese dataset Weibo21 and an accuracy of 87.8% in the GossipCop English dataset from FakeNewsNet. When ARG is combined with the LLM Judgement predictor, it achieves an accuracy of 77.4% and 88% in the Chinese and English datasets retrospectively. Finally, ARG-D achieves 77.2% and 87% accuracy scores in the Chinese and English datasets, respectively.

The SheepDog method [76], which extracts content-focused veracity attributions from LLMs, has been tested on the Politifact dataset and achieved an accuracy of 88.44% and on the GossipCop dataset and got an accuracy of 75.77%. On the LUN dataset, the SheepDog method achieved its highest accuracy (93.05%). In all datasets, SheepDog outperformed all other methods that rely on pre-trained models like BERT, RoBERTa, DeBerta, GPT3.5, or InstructGPT.

The Diverse reaction generation Explainable proxy tasks and LLM-Based expert ensemble method, also known as DELL [77] has been tested on seven datasets related to fake news detection (Pheme and llm-mis), framing detection and propaganda tactic detection. DELL achieved state-of-the-art performance against baselines in all benchmarks. This shows the success of integrating LLMs in multiple stages of news veracity evaluation.

5.3.4. Fake Profile Detection Using Language Patterns and Anomalies in Profile Content Analysis

In a different context than the previous works, the authors in [85] introduced MedGraph a multi-head attention-based graph neural network (GNN) for detecting malicious behaviors (friendship edge creation) on information exchange networks. The experiments on the UCI message and the Digg datasets show that MedGraph outperformed all other baselines (i.e., Node2Vec, GraphSAGE, DeepWalk, etc.) with an AUC of 0.85 in the UCI message graph and 0.96 in the Digg graph.

5.3.5. Multi-Modal Approaches for Fake Profile Detection

The authors in [41] developed a dataset of 1800 legitimate LinkedIn profiles (LLPs), 600 fake LinkedIn profiles (FLPs) and 1200 profiles generated by ChatGPT (CLPs). They used word embeddings and trained a classifier to evaluate the set of LLM-generated user profiles. Their proposed method, SSTE achieved an accuracy of 72.26% when using BERT and 76.12% when using RoBERTa. They also used CLPs as fake profiles in the training phase to identify FLPs in the testing phase and achieved 65.39% accuracy with BERT embeddings and 60.38% accuracy with RoBERTa. Finally, when the task was to discriminate between LLPs and FLPs the accuracy of SSTE was 96.33% and 95% with BERT and ROBERTa, respectively.

The approach proposed in [86] used pre-trained BERT in combination with Logistic Regression for classifying profiles as fake or real. They considered a rich set of impactful features: attribute-based, network-based, and user features and used LLMs to extract a few more features such as gender. The method achieved an accuracy score of 94.65% with the additional extracted features. In contrast, the accuracy with the original features was only 87.32%.

Table 5 provides a summary of all the traditional and LLM-based methods presented in the previous subsections in the two tasks of detecting Fake user profiles and Fake content.

Table 5. Comparison of Traditional Methods and LLMs for Fake Profiles and Fake News.

	Fake Profiles	**Fake News**
Traditional Methods	**Techniques:** Profile attribute analysis, Network analysis, Behavior, interaction patterns **Methods:** SVM-NN [36], Hybrid Model [40], Probabilistic [19,37],EGSLA [43], PetriNets [44], Boost and bagging [47], GWO [48], OneR [49] **Algorithms:** SVM-NN, Naive Bayes, EGSLA, PetriNets, AdaBoost, GWO, J48, Random Forest, REP Tree, JRip, OneR, Naive Bayes	**Techniques:** Keyword Analysis, Heuristic Rules **Methods:** KAFN-PULP [23], Attention-based ANN with multiple sources [24], IKA [25], Probabilistic [26] **Algorithms:** GNEE, Attention-based ANN, IKA, Naive Bayes classifier
LLMs	**Techniques:** Profile content analysis, Multi-modal approaches **Methods:** MedGraph [85], SSTE [41], Logistic Regression [86], Multi-head attention-based GNN [85] **Models:** BERT, RoBERTa	**Techniques:** Classification, Fact-checking and verification, Contextual analysis, Multi-modal content analysis **Methods:** [69], BiLSTM and Att-BiLSTM,[70], BERT with socio-contextual information [71], FakeNewsGPT4 [72], FFRR [73], HiSS [68], LLMs for weak labels [74], ARG [75], SheepDog [76], DELL [77] **Models:** Fine-tuned ChatGPT and Google Gemini, BERT, FakeNewsGPT4, GPT-3.5-Turbo, OpenAssistant-30B-FULL

5.4. Datasets

Table 6 summarizes the main datasets found in the literature and used by the studies analyzed in the previous sub-sections.

Table 6. Datasets Overview.

Task	**Dataset**	**Source**	**URL (accessed on 15 August 2024)**
Fake News	FakeNewsCorpus	Articles	https://github.com/several27/FakeNewsCorpus
	Twitter, Weibo	Social Media Posts	https://ieee-dataport.org/documents/weibo-and-twitter
	Weibo21	Social Media Posts	https://github.com/kennqiang/mdfend-weibo21
	En-3 (FakeNewsNet, COVID)	Articles, Social Media Posts	https://github.com/bigheiniu/MM-COVID, https://github.com/KaiDMML/FakeNewsNet
	BuzzFeed News	Articles	https://www.kaggle.com/datasets/mrisdal/fact-checking-facebook-politics-pages
	LIAR benchmark	Articles	https://paperswithcode.com/dataset/liar
	ArCOV19-Rumors	Social Media Posts	https://gitlab.com/bigirqu/ArCOV-19
	AraNews	Articles	https://github.com/UBC-NLP/wanlp2020_arabic_fake_news_detection
	CMU-MisCOV19	Social Media Posts	https://zenodo.org/records/4024154
	DGM	Social Media Posts	https://github.com/rshaojimmy/MultiModal-DeepFake
	RAWFC, LIAR-RAW	Articles	https://github.com/Nicozwy/CofCED/tree/main/Datasets
	EuvsDisinfo	Articles	https://github.com/FloFloB/Euvsdisinfo-dataset
	FA-KES	Articles	https://www.kaggle.com/datasets/mohamadalhasan/a-fake-news-dataset-around-the-syrian-war
	FakeNewsNet	Social Media Posts	https://github.com/KaiDMML/FakeNewsNet
	Politifact	Articles, Social Media Posts	https://www.kaggle.com/datasets/shivkumarganesh/politifact-factcheck-data
	Pheme	Social Media Posts	https://www.kaggle.com/datasets/usharengaraju/pheme-dataset
	llm-mis	Articles, Social Media Posts	https://github.com/llm-misinformation/llm-misinformation/

Table 6. *Cont.*

Task	Dataset	Source	URL (accessed on 15 August 2024)
Fake Profile	UCI message	Online Messages	[87]
	Digg	Social Media Posts	https://networkrepository.com/digg.php
	MIB	Social Media Profiles	http://mib.projects.iit.cnr.it/dataset.html
	Weibo	Social	https://www.kaggle.com/datasets/bitandatom/social-network-fake-account-dataset,
	freelancer.in	Job Portal Profiles	https://github.com/harshitkgupta/Fake-Profile-Detection-using-ML

6. Case Studies and Practical Applications

Various instances of using Large Language Models (LLMs) to combat fake news are available across different platforms, showing their practical uses and how effective they are. The LLMs deployed on social media like Facebook and Twitter help in analyzing the user-generated content in real-time. These models find Fake information by scanning posts, comments, and shared links. Social media platforms can use natural language processing and machine learning techniques to spot fake news patterns such as sensational phrases or unverified sources thus helping them take quick action faster than before hence reducing the spread of misinformation [88].

Also, this is advantageous to news aggregators since they help in keeping their information correct and pure. For instance, researchers use LLMs to eradicate fake news by checking their facts against multiple sources and ascertaining that the news articles are genuine. The develop tools that help to curate accurate and reliable news for readers and maintain their credibility. Besides, e-commerce platforms like Amazon (https://www.aboutamazon.com/news/amazon-ai/amazon-improves-customer-reviews-with-generative-ai, accessed on 15 August 2024) and eBay (https://innovation.ebayinc.com/tech/engineering/how-ebay-created-a-language-model-with-three-billion-item-titles/, accessed on 15 August 2024) have begun applying LLMs in order to monitor product reviews and descriptions. They can improve customer trust and make online shopping more secure by detecting fraudulent reviews or wrong information about products.

The evaluation results of existing implementations have been promising. For example, performance comparisons between rule-based traditional detection systems and LLM-based ones indicated that the latter have higher accuracy of detection for fake news. The spread of misinformation on social media platforms can be reduced using language model machine learning techniques, an improvement from previous methods. User impact and feedback also highlight the effectiveness of LLMs in real-world applications. Users have reported higher trust in platforms that actively utilize LLMs for fake news detection, appreciating the increased reliability of the information provided [89].

In short, their practical use against misinformation across social media platforms, news aggregator sites, and online marketplaces shows how beneficial large language models can be in reducing falsehoods. This examination points out a better performance by LLMs as compared to traditional methods with significant improvements in terms of precision and user confidence rates. The continuous development and more extensive application of LLMs are critical as we continue to face up to fake news and look forward to digital platforms where dependable information is shared among users.

7. Challenges and Limitations

The use of LLMs for fake news detection comes with many challenges. These challenges arise from both the complexities of the technology and the tricky nature of fake news. In the following, we examine the various limitations of LLMs, which include issues with data quality and bias, difficulties in understanding context, and the high computational

power needed. We also discuss the problems that arise when these models have to scale in order to handle large amounts of data, the need for better interpretability, and the ethical and privacy concerns. Furthermore, we explore how these models struggle to adapt to new types of fake content. By examining these aspects in detail, we aim to provide a clear understanding of what needs to be improved to make LLMs more effective in detecting fake news.

Data Quality and Bias: LLMs are trained on large datasets that may contain biases reflecting the prejudices and inaccuracies present in the source data. These biases can influence the model's outputs, making it difficult to reliably detect fake news. Certain topics, events, or viewpoints may be overrepresented or underrepresented in the training data, leading to skewed detection capabilities [4].

Contextual Understanding: Fake news often involves nuanced language and contextual subtleties that are challenging for LLMs to fully comprehend. The lack of deep understanding of the context can lead to misclassification. Ambiguous language and the use of irony, sarcasm, or satire in fake news can confuse LLMs, which might struggle to differentiate between genuine and deceptive content [90].

Computational Efficiency: LLMs are computationally intensive. They require substantial amounts of processing power and memory to function effectively. This high computational demand poses several challenges. The sheer volume of computations needed for inference makes it difficult to deploy these models on devices with limited resources, such as mobile phones or edge devices. Furthermore, running LLMs at scale consumes significant energy, which is both costly and environmentally unsustainable. This high energy consumption can limit the practical application of LLMs in real-time fake news detection. Additionally, ensuring low latency responses is critical for real-time applications like fake news detection. However, the computational complexity of LLMs can introduce delays, reducing their effectiveness in time-sensitive scenarios [91].

Scalability: Scalability is a major challenge when deploying LLMs for fake news detection across vast and diverse data sources. Processing massive volumes of text data from social media, news outlets, and other online sources requires a highly scalable infrastructure. LLMs struggle to efficiently scale up to handle this volume without significant computational overhead. Implementing LLMs in distributed systems to manage scalability introduces complexities in synchronization, consistency, and data distribution, which can affect the model's performance and reliability. The financial cost of scaling up LLMs is considerable. The need for powerful hardware and continuous model retraining makes it expensive to maintain large-scale deployments, limiting accessibility for smaller organizations or projects [74].

Interpretability: Interpretability refers to the ability to understand and explain how LLMs arrive at their decisions. This is particularly important in sensitive applications like fake news detection. LLMs are often described as "black boxes" because their internal decision-making processes are opaque. This lack of transparency makes it difficult to understand why a particular piece of news is classified as fake, reducing trust in the model's outputs. Providing clear and understandable explanations for the model's decisions is challenging. Without explainability, it is hard for users to verify the accuracy and fairness of the model's predictions. In many jurisdictions, regulations require that automated decision-making systems be explainable. The interpretability challenges of LLMs make it difficult to comply with these regulations, posing legal and ethical risks [92].

Adaptability to New Types of Fake Content: Fake news constantly evolves, with new tactics and forms emerging to evade detection. LLMs face significant challenges in adapting to these changes. LLMs are typically trained on large datasets that do not update in real time. This static nature means they may not be equipped to recognize and respond to new types of fake content that arise after the training period. Implementing continuous learning mechanisms to keep LLMs up-to-date with the latest fake news tactics is complex and resource-intensive. Without this, the models can become outdated quickly. While LLMs are powerful, they can struggle to generalize from known types of fake

news to novel forms. The subtle nuances and evolving strategies used in fake content require constant adaptation, which is challenging for models primarily designed for static learning [72].

Ethical and Privacy Concerns: The deployment of Large Language Models (LLMs) in fake news detection brings significant ethical and privacy challenges that must be meticulously addressed. LLMs require vast amounts of data for training, often including personal and sensitive information. Ensuring these data are protected and handled in compliance with regulations like the General Data Protection Regulation (GDPR) is critical to safeguard individual privacy rights. The potential for data breaches or misuse poses substantial risks, necessitating robust data security measures and transparency in data handling practices [89]. Besides, the ethical landscape of LLM usage is complex, as these models themselves can be weaponized to generate and disseminate misinformation, exacerbating the very issue they are intended to combat. This dual-use nature of LLMs underscores the need for stringent ethical guidelines and controls to prevent misuse. Ethical considerations also extend to ensuring fairness and avoiding biases within the models, which can inadvertently perpetuate harmful stereotypes or misinformation [93]. A responsible approach to LLM deployment involves ongoing ethical oversight, the establishment of clear accountability frameworks, and the implementation of rigorous standards to guide the ethical use of these powerful technologies. By addressing these privacy and ethical concerns, the integration of LLMs in fake news detection can be managed more responsibly, ensuring they contribute positively to the integrity of information in the digital age [94].

8. Conclusions and Future Research Directions

Large language models (LLMs) have a dual role in the fake news world, though they can be misused to produce extremely believable false news articles and create deceptive online profiles. However, their potential as a tool for detection cannot be ignored. Using enhanced natural language processing skills, LLMs are capable of examining massive amounts of text information, detecting inconsistencies and identifying patterns that are indicative of misinformation.

There is no doubt that LLMs' for fake news detection have limitations and challenges that need to be overcome in order to increase their effectiveness and trustworthiness. One of the most promising areas to focus on is computational modeling. This can be conducted by using a combination of traditional methods and modern deep learning approaches such as hybrid techniques. As an example, we can improve detection accuracy by fusing rule-based systems with neural networks to allow for more detailed analyses of fake content. Furthermore, transformer models, which are among the latest developments in deep learning architectures, offer new possibilities for refining and optimizing LLM performance so they become better equipped to deal with issues relating to fake news complexity.

The inclusion of multimodal data into the fight against fake news presents another huge challenge. Most existing LLM models are centered on text, but fake news is usually composed of other forms of media such as images and videos. We can obtain a better understanding of the content by developing approaches that would allow us to carry out analysis and correlation across different modalities using models. In this way, LLMs will be able to detect more inconsistencies and lies than when they take into account a single modality only. To create more powerful systems to counteract complex strategies employed by contemporary disinformation; it will be necessary to combine textual, video and image analyses.

Another thing that is important in the fight against fake news is working with domain experts. This partnership can help to improve the accuracy of detecting false information through better fact-checking. Additionally, journalists and people from other fields who are involved in producing LLMs can work together to ensure their models are more accurate and take into account context. By doing so, they will be able to create datasets that capture the richness of real-world situations, which would reduce bias and make LLMs more

reliable in general. Additionally, this kind of cooperation encourages cross-disciplinary study that combines expertise from fields like computer science, linguistics, psychology and media studies to develop holistic approaches for fighting misinformation.

The use of more durable and transparent models is very significant in terms of improving trust and usability in fake news detection technologies. With the application of explainable AI techniques, it will be possible to understand the reasons behind the models' decisions thereby addressing the 'black box' problem and enhancing interpretability. Such a level of openness on a compliance basis will enable users to have confidence in these tools. Moreover, simplified tools for detecting false information that is understandable by laypersons can be instrumental in empowering wider population groups. As such, we can enhance LLMs' efficiency and reliability as well as their positive contribution towards ensuring the truthfulness of information in this digital era if we center on these future developments.

In conclusion, this survey paper has highlighted the state of the art in the use of LLMs in the field of fake news, shedding light on both fake profiles and fake content and using the LLMs both as generators and as detectors. It also discussed the challenges and limitations associated with using Large Language Models (LLMs) for fake news detection, including issues related to data quality and bias, contextual understanding, computational efficiency, scalability, interpretability, adaptability, and ethical and privacy concerns. Nevertheless, the potential of LLMs to enhance fake news detection is significant as long as they are addressed by future developments. Therefore, continuing to investigate this issue is paramount so that more accurate models that are faster and open can be created. Secondly, interdisciplinary collaboration should be fostered in order to encourage innovation which will result in holistic solutions blending expertise from different fields. We, therefore, ask researchers, technologists, as well as experts in various fields to join hands in refining LLMs so that their responsible deployment can be ensured thus reinforcing collective effort in fighting against false information that stains the digital landscape's reputation.

Author Contributions: Conceptualization, I.V.; Data curation, C.C.; Formal analysis, Y.H.; Funding acquisition, I.V.; Investigation, E.P. and C.C.; Methodology, C.C. and Y.H.; Project administration, I.V.; Resources, E.P.; Validation, I.V.; Visualization, C.C.; Writing—original draft, E.P.; Writing—review and editing, I.V. and Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: The project has been partially funded by the National Program "Development of Digital Products and Services", of Action 16706 "Digital Transformation of Small and Medium Enterprises", of the National Recovery and Resilience Plan Greece 2.0, of the Information Society S.A., Greece (Project No. 107181), on behalf of the company S. Charitakis & Co. L.P.

Data Availability Statement: The study does not use any data. All the mentioned publicly available datasets are given with URLs and dates of last visit.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Keller, F.B.; Schoch, D.; Stier, S.; Yang, J. Political astroturfing on Twitter: How to coordinate a disinformation campaign. *Political Commun.* **2020**, *37*, 256–280. [CrossRef]
2. Bahad, P.; Saxena, P.; Kamal, R. Fake news detection using bi-directional LSTM-recurrent neural network. *Procedia Comput. Sci.* **2019**, *165*, 74–82. [CrossRef]
3. Nasir, J.A.; Khan, O.S.; Varlamis, I. Fake news detection: A hybrid CNN-RNN based deep learning approach. *Int. J. Inf. Manag. Data Insights* **2021**, *1*, 100007. [CrossRef]
4. Roy, P.K.; Chahar, S. Fake profile detection on social networking websites: A comprehensive review. *IEEE Trans. Artif. Intell.* **2020**, *1*, 271–285. [CrossRef]
5. Varlamis, I.; Michail, D.; Glykou, F.; Tsantilas, P. A survey on the use of graph convolutional networks for combating fake news. *Future Internet* **2022**, *14*, 70. [CrossRef]
6. Sun, Y.; He, J.; Cui, L.; Lei, S.; Lu, C.T. Exploring the Deceptive Power of LLM-Generated Fake News: A Study of Real-World Detection Challenges. *arXiv* **2024**, arXiv:2403.18249.

7. Kareem, W.; Abbas, N. Fighting lies with intelligence: Using large language models and chain of thoughts technique to combat fake news. In Proceedings of the International Conference on Innovative Techniques and Applications of Artificial Intelligence, León, Spain, 14–17 June 2013; Springer: Berlin/Heidelberg, Germany, 2023; pp. 253–258.
8. Li, X.; Zhang, Y.; Malthouse, E.C. Large Language Model Agent for Fake News Detection. *arXiv* **2024**, arXiv:2405.01593.
9. Vykopal, I.; Pikuliak, M.; Srba, I.; Moro, R.; Macko, D.; Bielikova, M. Disinformation capabilities of large language models. *arXiv* **2024**, arXiv:2311.08838.
10. Jiang, B.; Tan, Z.; Nirmal, A.; Liu, H. Disinformation detection: An evolving challenge in the age of llms. In Proceedings of the 2024 SIAM International Conference on Data Mining (SDM), Houston, TX, USA, 18–20 April 2024; pp. 427–435.
11. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
12. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
13. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
14. Farhangian, F.; Cruz, R.M.; Cavalcanti, G.D. Fake news detection: Taxonomy and comparative study. *Inf. Fusion* **2024**, *103*, 102140. [CrossRef]
15. Aïmeur, E.; Amri, S.; Brassard, G. Fake news, disinformation and misinformation in social media: A review. *Soc. Netw. Anal. Min.* **2023**, *13*, 30. [CrossRef] [PubMed]
16. Goldstein, J.A.; Sastry, G.; Musser, M.; DiResta, R.; Gentzel, M.; Sedova, K. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv* **2023**, arXiv:2301.04246.
17. Chen, C.; Shu, K. Can LLM-Generated Misinformation Be Detected? In Proceedings of the 12th International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024.
18. Joshi, S.; Nagariya, H.G.; Dhanotiya, N.; Jain, S. Identifying fake profile in online social network: An overview and survey. In *Proceedings of the Machine Learning, Image Processing, Network Security and Data Sciences: Second International Conference (MIND 2020), Silchar, India, 30–31 July 2020*; Proceedings, Part I 2; Springer: Berlin/Heidelberg, Germany, 2020; pp. 17–28.
19. Ersahin, B.; Aktaş, Ö.; Kılınç, D.; Akyol, C. Twitter fake account detection. In Proceedings of the 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, Turkey, 5–8 October 2017; pp. 388–392.
20. Su, J.; Cardie, C.; Nakov, P. Adapting Fake News Detection to the Era of Large Language Models. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, 16–21 June 2024; pp. 1473–1490.
21. Kumar, S.; West, R.; Leskovec, J. Disinformation on the web: Impact, characteristics, and detection of Wikipedia hoaxes. In Proceedings of the 25th International Conference on World Wide Web, Montreal, QC, Canada, 11–15 April 2016; pp. 591–602.
22. Ruffo, G.; Semeraro, A.; Giachanou, A.; Rosso, P. Studying fake news spreading, polarisation dynamics, and manipulation by bots: A tale of networks and language. *Comput. Sci. Rev.* **2023**, *47*, 100531. [CrossRef]
23. de Souza, M.C.; Gôlo, M.P.S.; Jorge, A.M.G.; de Amorim, E.C.F.; Campos, R.N.T.; Marcacini, R.M.; Rezende, S.O. Keywords attention for fake news detection using few positive labels. *Inf. Sci.* **2024**, *663*, 120300. [CrossRef]
24. da Silva, F.C.D.; Vieira, R.; Garcia, A.C.B. Can Machines Learn to Detect Fake News? A Survey Focused on Social Media. In Proceedings of the Hawaii International Conference on System Sciences, Maui, HI, USA, 8–11 January 2019.
25. Guo, H.; Zeng, W.; Tang, J.; Zhao, X. Interpretable Fake News Detection with Graph Evidence. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, Birmingham, UK, 25 October 2023; pp. 659–668.
26. Granik, M.; Mesyura, V. Fake news detection using naive Bayes classifier. In Proceedings of the 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kyiv, Ukraine, 29 May–2 June 2017; pp. 900–903.
27. Michail, D.; Kanakaris, N.; Varlamis, I. Detection of fake news campaigns using graph convolutional networks. *Int. J. Inf. Manag. Data Insights* **2022**, *2*, 100104. [CrossRef]
28. Vlachos, A.; Riedel, S. Fact checking: Task definition and dataset construction. In Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, Baltimore, MD, USA, 26 June 2014; pp. 18–22.
29. PolitiFact. Available online: https://www.politifact.com/ (accessed on 28 July 2024).
30. Hu, L.; Wei, S.; Zhao, Z.; Wu, B. Deep learning for fake news detection: A comprehensive survey. *AI Open* **2022**, *3*, 133–155. [CrossRef]
31. Murayama, T. Dataset of fake news detection and fact verification: A survey. *arXiv* **2021** arXiv:2111.03299.
32. Snopes. Available online: https://www.snopes.com/ (accessed on 28 July 2024).
33. Suggest. Available online: https://www.suggest.com/ (accessed on 28 July 2024).
34. FastCheck. Available online: https://www.factcheck.org/ (accessed on 28 July 2024).
35. Fullfact. Available online: https://fullfact.org/ (accessed on 28 July 2024).
36. Khaled, S.; El-Tazi, N.; Mokhtar, H.M. Detecting fake accounts on social media. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 3672–3681.
37. Adikari, S.; Dutta, K. Identifying fake profiles in Linkedin. *arXiv* **2020**, arXiv:2006.01381.
38. Shah, A.; Varshney, S.; Mehrotra, M. Detection of Fake Profiles on Online Social Network Platforms: Performance Evaluation of Artificial Intelligence Techniques. *SN Comput. Sci.* **2024**, *5*, 1–15. [CrossRef]

39. Bertini, F.; Sharma, R.; Montesi, D. Are social networks watermarking us or are we (unawarely) watermarking ourself? *J. Imaging* **2022**, *8*, 132. [CrossRef]
40. Smruthi, M.; Harini, N. A hybrid scheme for detecting fake accounts in Facebook. *Int. J. Recent Technol. Eng. (IJRTE)* **2019**, *7*, 2277–3878.
41. Ayoobi, N.; Shahriar, S.; Mukherjee, A. The looming threat of fake and LLM-generated linkedin profiles: Challenges and opportunities for detection and prevention. In Proceedings of the 34th ACM Conference on Hypertext and Social Media, Rome, Italy, 4–8 September 2023; pp. 1–10.
42. Sudhakar, T.; Gogineni, B.C.; Vijaya, J. Fake profile identification using machine learning. In Proceedings of the 2022 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE), Naya Raipur, India, 30–31 December 2022; pp. 47–52.
43. Balaanand, M.; Karthikeyan, N.; Karthik, S.; Varatharajan, R.; Manogaran, G.; Sivaparthipan, C. An enhanced graph-based semi-supervised learning algorithm to detect fake users on Twitter. *J. Supercomput.* **2019**, *75*, 6085–6105. [CrossRef]
44. Sahoo, S.R.; Gupta, B.B. Hybrid approach for detection of malicious profiles in Twitter. *Comput. Electr. Eng.* **2019**, *76*, 65–81. [CrossRef]
45. Yang, Z.; Wilson, C.; Wang, X.; Gao, T.; Zhao, B.Y.; Dai, Y. Uncovering social network sybils in the wild. *ACM Trans. Knowl. Discov. Data (TKDD)* **2014**, *8*, 1–29. [CrossRef]
46. Mewada, A.; Dewang, R.K. CIPF: Identifying fake profiles on social media using a CNN-based communal influence propagation framework. *Multimed. Tools Appl.* **2024**, *83*, 29419–29454. [CrossRef]
47. Patil, D.R.; Pattewar, T.M.; Punjabi, V.D.; Pardeshi, S.M. Detecting Fake Social Media Profiles Using the Majority Voting Approach. *EAI Endorsed Trans. Scalable Inf. Syst.* **2024**, *11*. [CrossRef]
48. Sekkal, N.; Mahammed, N. Grey Wolf Optimizer Algorithm To Detection Fake Profile In Facebook. In Proceedings of the 2023 16th International Conference on Developments in eSystems Engineering (DeSE), Istanbul, Turkiye, 18–20 December 2023; pp. 402–406.
49. Gupta, A.; Kaushal, R. Towards detecting fake user accounts in Facebook. In Proceedings of the 2017 ISEA Asia Security and Privacy (ISEASP), Surat, India, 29 January–1 February 2017; pp. 1–6.
50. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef]
51. Pires, T.; Schlinger, E.; Garrette, D. How Multilingual is Multilingual BERT? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 4996–5001.
52. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
53. Fn-bert. 2023. Available online: https://huggingface.co/ungjus/Fake_News_BERT_Classifier (accessed on 28 July 2024).
54. Zellers, R.; Holtzman, A.; Rashkin, H.; Bisk, Y.; Farhadi, A.; Roesner, F.; Choi, Y. Defending against neural fake news. In Proceedings of the Advances in Neural Information Processing Systems 32 (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019.
55. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv* **2023**, arXiv:2307.09288.
56. Abdin, M.; Jacobs, S.A.; Awan, A.A.; Aneja, J.; Awadallah, A.; Awadalla, H.; Bach, N.; Bahree, A.; Bakhtiari, A.; Behl, H.; et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv* **2024**, arXiv:2404.14219.
57. OpenAI. Chatgpt 3.5. Available online: https://chat.openai.com/chat (accessed on 28 July 2024).
58. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv* **2023**, arXiv:2303.08774.
59. Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; et al. Qwen2 technical report. *arXiv* **2024**, arXiv:2407.10671
60. Jiang, A.Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D.S.; Casas, D.d.l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. Mistral 7B. *arXiv* **2023**, arXiv:2310.06825.
61. Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A.M.; Hauth, A.; et al. Gemini: A family of highly capable multimodal models. *arXiv* **2023**, arXiv:2312.11805.
62. Bhat, M.M.; Parthasarathy, S. How effectively can machines defend against machine-generated fake news? an empirical study. In Proceedings of the 1st Workshop on Insights from Negative Results in NLP, Online, 19 November 2020; pp. 48–53.
63. Wang, Z.; Cheng, J.; Cui, C.; Yu, C. Implementing BERT and fine-tuned RobertA to detect AI generated news by ChatGPT. *arXiv* **2023**, arXiv:2306.07401.
64. Sun, Y.; He, J.; Lei, S.; Cui, L.; Lu, C.T. Med-mmhl: A multi-modal dataset for detecting human-and llm-generated misinformation in the medical domain. *arXiv* **2023**, arXiv:2306.08871.
65. Pan, Y.; Pan, L.; Chen, W.; Nakov, P.; Kan, M.Y.; Wang, W. On the Risk of Misinformation Pollution with Large Language Models. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, 6–10 December 2023; pp. 1389–1403.

66. Popat, K.; Mukherjee, S.; Yates, A.; Weikum, G. DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 22–32.
67. Shu, K.; Cui, L.; Wang, S.; Lee, D.; Liu, H. defend: Explainable fake news detection. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 395–405.
68. Zhang, X.; Gao, W. Towards LLM-based Fact Verification on News Claims with a Hierarchical Step-by-Step Prompting Method. In Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Nusa Dua, Indonesia, 1–4 November 2023; pp. 996–1011.
69. Boissonneault, D.; Hensen, E. Fake News Detection with Large Language Models on the LIAR Dataset. *Research Square* **2024**. [CrossRef]
70. AlEsawi, B.; Al-Tai, M.H. Detecting Arabic Misinformation Using an Attention Mechanism-Based Model. *Iraqi J. Comput. Sci. Math.* **2024**, *5*, 285–298. [CrossRef]
71. Upadhyay, R.; Pasi, G.; Viviani, M. Leveraging Socio-contextual Information in BERT for Fake Health News Detection in Social Media. In Proceedings of the 3rd International Workshop on Open Challenges in Online Social Networks, Rome, Italy, 4–8 September 2023; pp. 38–46.
72. Liu, X.; Li, P.; Huang, H.; Li, Z.; Cui, X.; Liang, J.; Qin, L.; Deng, W.; He, Z. FakeNewsGPT4: Advancing Multimodal Fake News Detection through Knowledge-Augmented LVLMs. *arXiv* **2024**, arXiv:2403.01988.
73. Zhang, X.; Gao, W. Reinforcement Retrieval Leveraging Fine-grained Feedback for Fact Checking News Claims with Black-Box LLM. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italy, 20–25 May 2024; pp. 13861–13873.
74. Leite, J.A.; Razuvayevskaya, O.; Bontcheva, K.; Scarton, C. Detecting misinformation with llm-predicted credibility signals and weak supervision. *arXiv* **2023**, arXiv:2309.07601.
75. Hu, B.; Sheng, Q.; Cao, J.; Shi, Y.; Li, Y.; Wang, D.; Qi, P. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 22105–22113.
76. Wu, J.; Hooi, B. Fake News in Sheep's Clothing: Robust Fake News Detection Against LLM-Empowered Style Attacks. *arXiv* **2023**, arXiv:2310.10830.
77. Wan, H.; Feng, S.; Tan, Z.; Wang, H.; Tsvetkov, Y.; Luo, M. DELL: Generating reactions and explanations for llm-based misinformation detection. *arXiv* **2024**, arXiv:2402.10426.
78. Whitehouse, C.; Weyde, T.; Madhyastha, P.; Komninos, N. Evaluation of fake news detection with knowledge-enhanced language models. In Proceedings of the International AAAI Conference on Web and Social Media, Atlanta, GA, USA 6–9 June 2022; Volume 16, pp. 1425–1429.
79. Aggarwal, A.; Chauhan, A.; Kumar, D.; Verma, S.; Mittal, M. Classification of fake news by fine-tuning deep bidirectional transformers based language model. *EAI Endorsed Trans. Scalable Inf. Syst.* **2020**, *7*, e10. [CrossRef]
80. Su, J.; Zhuo, T.Y.; Mansurov, J.; Wang, D.; Nakov, P. Fake news detectors are biased against texts generated by large language models. *arXiv* **2023**, arXiv:2309.08674.
81. Wang, B.; Ma, J.; Lin, H.; Yang, Z.; Yang, R.; Tian, Y.; Chang, Y. Explainable Fake News Detection With Large Language Model via Defense Among Competing Wisdom. In Proceedings of the ACM on Web Conference 2024, Singapore, 13–17 May 2024; pp. 2452–2463.
82. Dulhanty, C.; Deglint, J.L.; Daya, I.B.; Wong, A. Taking a stance on fake news: Towards automatic disinformation assessment via deep bidirectional transformer language models for stance detection. *arXiv* **2019**, arXiv:1911.11951.
83. Jin, R.; Fu, R.; Wen, Z.; Zhang, S.; Liu, Y.; Tao, J. Fake News Detection and Manipulation Reasoning via Large Vision-Language Models. *arXiv* **2024**, arXiv:2407.02042.
84. Aman, M. Large Language Model Based Fake News Detection. *Procedia Comput. Sci.* **2024**, *231*, 740–745. [CrossRef]
85. Chen, K.; Wang, Z.; Liu, K.; Zhang, X.; Luo, L. MedGraph: Malicious edge detection in temporal reciprocal graph via multi-head attention-based GNN. *Neural Comput. Appl.* **2023**, *35*, 8919–8935. [CrossRef]
86. Vyawahare, M.; Govilkar, S. Fake profile recognition using profanity and gender identification on online social networks. *Soc. Netw. Anal. Min.* **2022**, *12*, 170. [CrossRef]
87. Opsahl, T.; Panzarasa, P. Clustering in weighted networks. *Soc. Netw.* **2009**, *31*, 155–163. [CrossRef]
88. Tschiatschek, S.; Singla, A.; Gomez Rodriguez, M.; Merchant, A.; Krause, A. Fake news detection in social networks via crowd signals. In Proceedings of the Companion Proceedings of the the Web Conference 2018, Lyon, France, 23–27 April 2018; pp. 517–524.
89. Shah, S.B.; Thapa, S.; Acharya, A.; Rauniyar, K.; Poudel, S.; Jain, S.; Masood, A.; Naseem, U. Navigating the Web of Disinformation and Misinformation: Large Language Models as Double-Edged Swords. *IEEE Access* **2024**. [CrossRef]
90. Zhou, X.; Zafarani, R.; Shu, K.; Liu, H. Fake news: Fundamental theories, detection strategies and challenges. In Proceedings of the 12th ACM International Conference on Web Search and Data Mining, Melbourne, VIC, Australia, 11–15 February 2019; pp. 836–837.

91. Stojkovic, J.; Choukse, E.; Zhang, C.; Goiri, I.; Torrellas, J. Towards Greener LLMs: Bringing Energy-Efficiency to the Forefront of LLM Inference. *arXiv* **2024**, arXiv:2403.20306.
92. Capuano, N.; Fenza, G.; Loia, V.; Nota, F.D. Content-based fake news detection with machine and deep learning: A systematic review. *Neurocomputing* **2023**, *530*, 91–103. [CrossRef]
93. Sarker, I.H. LLM potentiality and awareness: A position paper from the perspective of trustworthy and responsible AI modeling. *Discov. Artif. Intell.* **2024**, *4*, 40. [CrossRef]
94. Barman, D.; Guo, Z.; Conlan, O. The dark side of language models: Exploring the potential of llms in multimedia disinformation generation and dissemination. *Mach. Learn. Appl.* **2024**, *16*, 100545. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.