electronics

MDPI

Article

Artificial Intelligence in Social Media Forensics: A Comprehensive Survey and Analysis

Biodoumoye George Bokolo * and Qingzhong Liu

Department of Computer Science, Sam Houston State University, Huntsville, TX 77341, USA; qxl005@shsu.edu
* Correspondence: bgb023@shsu.edu

Abstract: Social media platforms have completely revolutionized human communication and social interactions. Their positive impacts are simply undeniable. What has also become undeniable is the prevalence of harmful antisocial behaviors on these platforms. Cyberbullying, misinformation, hate speech, radicalization, and extremist propaganda have caused significant harms to society and its most vulnerable populations. Thus, the social media forensics field was born to enable investigators and law enforcement agents to better investigate and prosecute these cybercrimes. This paper surveys the latest research works in the field to explore how artificial intelligence (AI) techniques are being utilized in social media forensics investigations. We examine how natural language processing can be used to identify extremist ideologies, detect online bullying, and analyze deceptive profiles. Additionally, we explore the literature on GNNs and how they are applied in social network modeling for forensic purposes. We conclude by discussing the key challenges in the field and suggest future research directions.

Keywords: social media forensics; digital forensics; artificial intelligence; natural language processing; graph neural networks; Generative Adversarial Networks; computer science; computational social science; data intelligence; network forensics

Citation: Bokolo, B.G.; Liu, Q. Artificial Intelligence in Social Media Forensics: A Comprehensive Survey and Analysis. *Electronics* **2024**, *13*, 1671. https://doi.org/10.3390/electronics13091671

Academic Editors: Umit Karabiyik, Mamoun Alazab and Abdelkader Ouda

Received: 15 March 2024
Revised: 15 April 2024
Accepted: 20 April 2024
Published: 26 April 2024

1. Introduction

The adoption and usage of online social networks have grown exponentially over the years. In the eight years between 2015 and 2023, there has been a 138.2% increase in users of social media platforms, growing from 2.08 billion to 4.95 billion users [1]. This growth is hardly surprising as these platforms have revolutionized individual communication, and transformed collaboration and information dissemination in this digital age. Billions of worldwide users now engage daily in a myriad of online interactions. The massive amount of data generated from these interactions is one of the foremost sources of digital evidence used in social media forensics investigations like background checks or intelligence gathering [2]. This proves a vital resource, as the exponential growth seen in social media usage also correlates directly to a surge in cybercrimes [3]; cybercrime costs have risen from USD 3 trillion in 2015 to USD 8 trillion in 2023 [4]. This surge in cybercrimes prompted a paradigm shift in the field of digital forensics, necessitating research into novel methodologies and tools to effectively investigate and mitigate crimes and antisocial behavior within the social media landscape.

This transformative shift hinges on the integration of artificial intelligence (AI) technologies into conventional digital forensic practices. Two AI techniques, natural language processing and deep learning architectures, hold substantial promise for enhancing the capabilities of forensic investigators: they can automate labor-intensive tasks, unearth concealed patterns, and even extract actionable insights from diverse datasets that are an intrinsic part of social media platforms [5]. Nonetheless, we face challenges when harnessing these AI solutions. Integrating AI into social media forensics creates a complicated environment, privacy preservation concerns contradict data integrity issues, and algorithmic bias appears

in tandem with ethical quandaries, all requiring meticulous navigation and deliberation. This necessitates an exhaustive examination of the latest methodologies, advancements, and approaches in digital forensics combined with AI for social media analysis.

This research work tries to fill the void in the existing literature by undertaking a comprehensive survey of the symbiotic interplay between social media forensics and AI. Our objectives are twofold: to underscore the significance of AI and its potential in facilitating social media forensics investigations; and also to emphasize the practical considerations of ethics and data governance in the deployment of AI-driven forensics tools. Our work is divided into six sections. Section 2 explores digital forensics—the umbrella field for all forensics investigations that involve digital evidence. We discuss the different subdomains in digital forensics, highlighting the various existing literature in each subdomain. In Section 3, we delve into social media forensics—our primary research focus. We explore its challenges and complexities, as well as the conventional techniques that have hitherto been used in the field. In Sections 4–6, we examine the practical applications of AI in social media forensics—considering three leading techniques: natural language processing (NLP), graph neural networks (GNNs), and generative adversarial networks (GANs). We conclude in Section 7 by discussing the observed challenges in the field, and we suggest areas of potential future research.

By highlighting the opportunities and complexities inherent in the adoption of AI in social media forensics, we aim to enrich contemporary conversations about the responsible and effective use of AI solutions in the pursuit of safety and justice in this digital age. Ultimately, to guarantee the safety of the social media platforms we have all grown accustomed to it is crucial we understand how we can adopt AI and other emerging technologies to aid social media forensics.

2. Digital Forensics

In the ever-evolving digital landscape, where information and interactions increasingly reside within a virtual realm, the need for digital forensics has become paramount. For investigative and security purposes, the digital forensics field ensures that digital evidence is identified, preserved, acquired, analyzed, and presented efficiently [6]. Just like any other forensic investigation, the sole aim is to obtain evidence and obtain more knowledge about a particular incident [6,7]. Traditional forensic processes have historically focused on identifying chemical, biological, or physical traces; digital forensics, however, deals with new forms of traces [8,9]. The use of various technology devices, which have become widespread in modern societies, produces these new traces, that are described as "digital". When digital devices are used for criminal activities, they leave a digital trace (or footprint) that is then collected to serve as the evidentiary data for the forensics investigation process.

2.1. Foundation and Methodology

Digital forensics, unlike its traditional counterpart, delves into the intangible world of digital information. At its core lies the fundamental principle of preserving evidentiary integrity [10]. This entails maintaining the chain of custody and authenticity of the data (evidence) during the whole investigation, starting from identification and acquisition to analysis and presentation. This rigorous approach ensures that digital evidence presented in court or used for internal investigations remains legally admissible and trustworthy. To perpetually live up to this fundamental principle, the following rules are vital to the digital forensics investigation.

1. Admissibility: The goal of every action must be to preserve digital evidence in a way that makes it acceptable in court or other legal proceedings.
2. Chain of custody: A meticulous record must be maintained to demonstrate the origin and handling of evidence throughout the investigation, ensuring its authenticity and integrity.
3. Minimization of modification: Whenever possible, data should be acquired in a way that minimizes or prevents modifications to the original evidence.

4. Documentation: Every step of the investigation process must be comprehensively documented, including tools used, procedures followed, and analysis performed.
5. Validation: All analytical methods and tools used must be scientifically sound and validated to ensure their reliability and repeatability.

In Ref. [1], a high-level digital forensics methodology is proposed by the National Institute of Standards and Technology (NIST). Figure 1 depicts the four phases of this methodology.

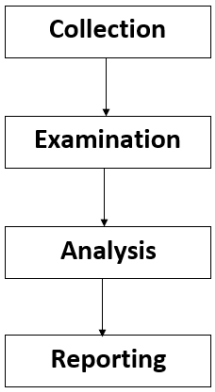

Figure 1. NIST digital forensics methodology.

1. Collection: The process of locating and recording credible sources of data pertinent to the incident, followed by the acquisition of data from these sources while ensuring their integrity is maintained.
2. Examination: The assessment of data obtained during the collection phase, focusing on extracting relevant information related to the incident while maintaining the integrity of the data.
3. Analysis: The study of information extracted during the examination phase to address pertinent investigative questions and to determine if a conclusive or partial conclusion can be reached.
4. Reporting: The preparation and presentation of the investigation's procedures, methodologies, and tools utilized, along with the outcomes derived from the analysis phase.

2.2. Domains in Digital Forensics

The ever-expanding digital landscape has given rise to diverse specialized areas within the overall field of digital forensics. In Figure 2, we illustrate the six major subdomains within digital forensics. Each domain requires specific expertise and methodologies due to the unique challenges and intricacies associated with their different data sources. While Roux et al. [11] argue that the field of forensics should not be split into subdomains, we can always identify at least five unique sources of evidential information (subdomains) that can be utilized in digital forensics investigations [12]. In subsequent sections, we will examine these five subdomains in digital forensics, with their data sourced from computers, mobile devices, network, databases, and the cloud. The conspicuous omission of social media as a data source in this list is due to the paper's emphasis on social media forensics in subsequent sections. Table 1 provides a summary of the reviewed papers in each of the examined five domains, and their applied methodologies.

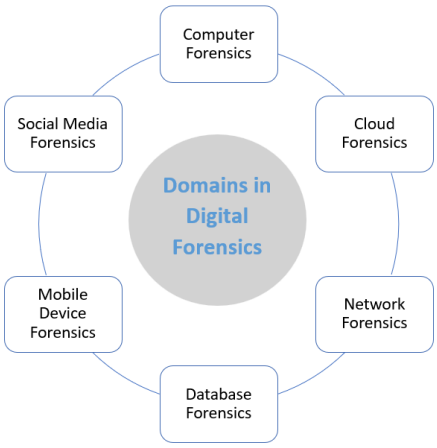

Figure 2. Domains in digital forensics.

Table 1. Overview of the subdomains in digital forensics.

Domain	Research Studies	Common Methodologies
Computer Forensics	[13–20]	Forensic data acquisition; file and operating system analysis; steganalysis
Mobile Device Forensics	[21,22]	Time synchronization; intra- and inter-application analysis; media log analysis; file system analysis
Network Forensics	[23–26]	Deep neural networks with PSO algorithm; network packet analysis; network log analysis
Database Forensics	[27–31]	Audit logs analysis; tamper detection; forensic data recovery
Cloud Forensics	[21,32–34]	Time synchronization; intra- and inter-application analysis; CLASS; forensic logging frameworks

2.3. Computer Forensics

Computer forensics is a multidisciplinary domain that combines computer science, sociology, and law with the goal of gathering and analyzing data from computer systems and their storage drives. The information gathered from these sources can then be used as evidence in legal proceedings [35]. Computer forensics is used in various criminal investigations, including but not limited to cybercrimes, identity theft, financial fraud, child pornography, homicide, and rape [13].

Computer forensics investigators employ an arsenal of techniques to extract and analyze evidence. These techniques include:

1. Data acquisition: Data acquisition in computer forensics is the process of gathering and recovering sensitive data during a digital forensic investigation [36,37]. This process involves capturing digital data from various sources such as disk, RAM, swap files, operating systems, and other storage mediums [13]. Overall, data acquisition is an important aspect of computer forensics, and it is essential for investigators to have the necessary skills and tools to identify and capture digital evidence effectively. In their 2021 paper, Ref. [14] demonstrated the importance of data acquisition and recovery in the computer forensics process. Using Photorec, they were able to recover 2781 files of different data formats that had been previously deleted from a 32 gigabyte flash drive [38]. Phorotec is a popular data acquisition and forensics tool used by computer forensics investigators. Other tools include FTK imager and TestDisk.

2. File system analysis: This is the data structure that makes it possible to store, access, and retrieve data efficiently on a computer system; without it, all files would become disorganized and tedious to access. File system analysis in the context of computer forensics involves examining the structure and contents of the file system, recovering deleted files, and reconstructing file activity timelines. Using file system activities, Khan et al. [15] presented a post-event timeline reconstruction method based on artificial neural network technology. By following earlier file system activities, they were able to map the chronology of important events on the computer system using a neural network methodology. Ref. [39] investigates and assesses the suitability of neural network approaches in computer forensics investigation by examining data associated with the file system of the computer to ascertain whether it has been altered by a particular application.
3. Operating system analysis: Operating system analysis involves finding and evaluating relevant data from the operating system of the concerned computer or digital system [40]. With an emphasis on forensic memory acquisition, Huebner et al. [16] discuss how operating system design and implementation affect computer forensics investigation methodology. The operating system might theoretically facilitate investigative inquiries by providing instruments for data analysis and by facilitating easy access to system data. Ref. [17] offers a thorough overview of the literature on operating system logs forensic analysis. Due to their ability to capture crucial system activity, these system event logs are among the foremost sources of digital evidence in forensic cases.
4. Steganography and data hiding detection: Steganography involves concealing information within a carrier, while steganalysis refers to the procedure of identifying concealed information within a carrier [18]. In May 2011, the German Federal Criminal Police (BKA) detained an Al-Qaeda affiliate in Berlin, seizing a chip holding a folder protected by a password. Through forensic analysis, specialists managed to decrypt the folder, exposing a pornographic video labeled 'KickAss', housing 141 discrete files outlining future targets and activities of Al-Qaeda [41]. Identifying concealed data embedded within files or unused sectors of storage devices, a technique frequently employed by criminals to hide sensitive information, has emerged as a critical area of research. Davidson et al.'s [19] research focused on developing a prototype software (version 1.0) that can detect if an image has any concealed or encrypted information in it. The software prototype was built using a sophisticated artificial neural network (ANN) system. Ref. [20] presents an innovative method of JPEG image steganalysis. This is driven by the need for a rapid and precise identification of concealed data and stego-carriers within image file datasets. As advances are being made in the field of steganalysis, malicious actors keen to stay ahead of the law are intensifying efforts to hide their data, potentially resorting to algorithms deliberately crafted to circumvent detection during forensic investigations.

2.4. Mobile Device Forensics

In today's hyper-connected society, smartphones and tablets have become more than just communication tools—they are repositories of our personal and professional lives, storing a wealth of information from messages and contacts to location data and financial transactions. This ubiquitous presence has also made them targets for cybercriminals and investigators alike, leading to the emergence of mobile device forensics as an important subdomain within the digital forensics field. Sharma et al. [21] present a mobile cloud forensic process that combines the conventional forensic procedure with time synchronization and intra- and inter-application analysis. Time synchronization is the process of ensuring that the time settings across different devices or systems in question are aligned. In the context of mobile device forensics, this is crucial because it allows investigators to correlate events accurately across various data sources such as mobile devices and cloud services. Since every forensic tool has its limitations, investigators must also be aware of

which tools possess the capability to manage mobile forensics of specific mobile devices and artefacts. Most mobile phones are not built for data security and privacy, making them a common communication monitoring device. The aim of Ref. [22] was to propose a mobile forensic workflow for detecting and analyzing embedded threats that could be used as a surveillance tool at various levels of a mobile device.

2.5. Network Forensics

Network forensics unravels the hidden threads of communication and activity within networks. This specialized subdomain of digital forensics delves into the analysis of network traffic, logs, and infrastructure, providing crucial insights into cyberattacks, data breaches, and other malicious activities. Unlike its computer and mobile device counterparts, network forensics does not directly deal with physical devices but rather the dynamic flow of data across networks. This dynamic nature presents unique challenges which include volatility, large data volumes, and encrypted traffic. To surmount these challenges, network forensics researchers employ a variety of techniques like packet capture, traffic analysis, log analysis, and network forensics tools. In Ref. [23], the author explores a novel approach to network forensics, outlining the stages of a digital investigation for locating and following attack patterns in Internet of Things (IoT) networks. Their suggested framework was optimized to perform three new tasks: (1) garnering network data-flows and verifying their integrity to manage secured networks; (2) automatically fine-tuning the deep learning variables; and (3) developing a deep neural network (DNN) model to detect and monitor abnormal occurrences within IoT networks in smart homes. Ref. [24] reviews how deep packet inspection and other packet analysis techniques are used in network forensics and offers an overview of AI-powered techniques for advanced network traffic classification and pattern recognition. The characteristics of both packet analyzer software and hardware appliances are examined from the standpoint of their possible application in network forensics.

2.6. Database Forensics

Databases, the hidden repositories of information powering organizations and online services, store a wealth of sensitive data. When investigations demand uncovering fraudulent activity, analyzing data breaches, or recovering deleted information, database forensics emerges as a specialized and crucial field in digital forensics. Database forensics is the area of digital forensic investigation that involves the analysis of databases, including the raw data and the metadata that describe the data [42]. Unlike other digital forensics areas, database forensics deals specifically with structured data organized within databases. This presents some unique challenges. Firstly, the diversity of data types necessitates specialized techniques tailored to each database format, whether relational or NoSQL. Secondly, navigating the complexity of database schemas and comprehending data relationships is essential for precise analysis and interpretation. Lastly, managing the volatility of transaction logs and audit logs, which are critical for tracing activity, demands swift and meticulous acquisition procedures. Reconstructing, gathering, preserving, analyzing, and reporting database incidents are all part of database forensic investigation (DBFI). Various forensic methodologies, principles, and undertakings have been examined in order to resolve select database case-studies using a limited number of DBFI process models [43]. Every interaction in a database may leave a digital trail, and the majority of database breaches are focused on compromising the three main security objectives (confidentiality, integrity, and authenticity) of the data contained within. Consequently, Ref. [43] suggest appropriate procedures for building an integrated incident response model (IIRM) that is dependable in the field of database forensic investigation.

2.7. Cloud Forensics

The ever-expanding digital landscape has brought about the advent of the cloud—an essential storage and computing platform for individuals and organizations alike. How-

ever, this vast and dynamic environment presents unique challenges for investigations, necessitating the emergence of cloud forensics. This subdomain of digital forensics delves into the complexities of cloud-based data and infrastructure, uncovering evidence and reconstructing activities within the digital cloud. Government agencies and researchers alike have extensively documented the difficulties associated with cloud forensics, even though a lot of the problems still remain unsolved. Supplementing the standard forensic procedure, Ref. [21] introduces a mobile cloud forensic methodology that includes time synchronization and intra- and inter-application analysis. An essential step in enabling the forensic analyst to quickly conduct their investigation in the mobile cloud is time synchronization. The investigation may be harmed by an anti-forensic attacker in the cloud who manipulates evidence and sways the cloud forensic process. Rani et al. [32] suggest effective algorithms for the safe transfer of data/evidence and the early identification of anti-forensic attacks (AFAs). The majority of secure logging solutions on the market today are made for conventional systems rather than the intricacies of cloud environments. To protect logs in a cloud environment, Ref. [33] suggests an alternate method called the Cloud Log Assuring Soundness and Secrecy (CLASS) process. The forensics investigation encounters novel legal obstacles when digital evidence is stored in a cloud storage environment. Ref. [44] identifies three primary legal issues brought on by the state of cloud computing today—possession (ownership of cloud content), territoriality (loss of location), and confiscation procedure (problems with user authentication and data preservation). Cloud forensics researchers developed a novel framework through a design science research methodological (DSRM) approach [34]. The findings of their case study show that their framework can help address the difficulties and complexities encountered by digital forensic investigators when gathering legally permissible digital evidence from the cloud environment.

3. Social Media Forensics Fundamentals

Online social networks have become ingrained in our daily lives, shaping communication, information consumption, and even influencing real-world events. Social media forensics involves investigating and analyzing digital information from these ubiquitous social networks to gather evidence for legal, investigative, or intelligence purposes. This evidence encompasses a diverse range of data types, including:

(a) Textual content: Posts, comments, messages, and other text-based interactions provide insights into user behavior, opinions, and potential criminal activities like cyberbullying or hate speech.
(b) Multimedia evidence: Images, videos, and even audio recordings can reveal crucial details about events, locations, and individuals involved in investigations.
(c) Network connections: Analyzing user connections, groups, and interactions can shed light on criminal networks, organized groups, or hidden associations.
(d) Metadata: Timestamps, location data, and other embedded information within social media content can provide valuable context and forensic clues.

Figure 3 provides an overview of the fundamental objective of social media forensics, which revolves around utilizing the diverse digital evidence to:

(a) Reconstruct past events: By meticulously piecing together user activity and interactions, investigators can reconstruct timelines, identify key players, and understand the context surrounding specific situations. This is crucial in criminal investigations or even analyzing the spread of misinformation.
(b) Identify criminal activity: Social media platforms, unfortunately, can be breeding grounds for illegal activities like cyberbullying, hate speech, online harassment, and even fraud. Forensic analysis can uncover evidence of these crimes, supporting legal proceedings and ensuring user accountability.
(c) Unveil hidden networks: Analyzing social media connections can reveal patterns of communication and association, aiding investigations into organized crime, terrorist

groups, or other criminal networks. This plays a vital role in disrupting illegal activities and ensuring public safety.

(d) Analyze public opinion and sentiment: By analyzing large datasets of social media posts, we can gain valuable insights into public opinion on various topics. This information empowers researchers, organizations, and even governments to understand societal trends and make informed decisions.

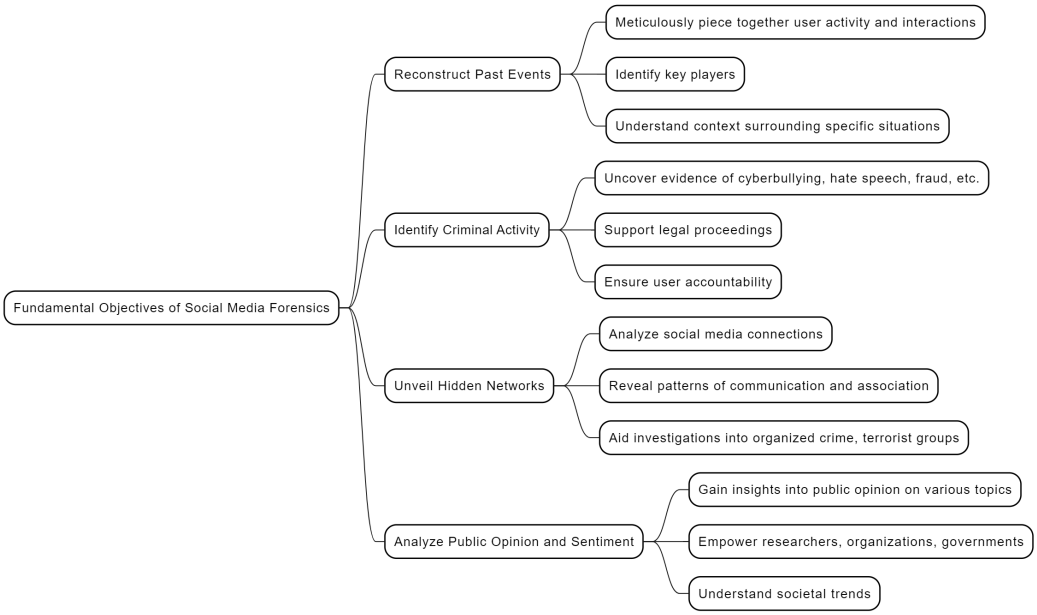

Figure 3. Social media forensics objectives at a glance.

3.1. Key Challenges and Complexities

As opposed to traditional digital forensics, social media forensics presents its own set of unique challenges and complexities for researchers. These challenges often arise from the unique nature of social media platforms and the digital evidence they contain. In this subsection, we discuss some of these challenges:

(a) Data volatility: The dynamic nature of digital information generated and shared on social media platforms means social media data can be highly transient, with content frequently changing or being deleted entirely. This volatility arises due to several factors: real-time updates, where users can post updates, comments, and messages instantly, leading to a continuous stream of new data; user control, allowing individuals to edit or delete their posts and comments, impacting the availability of data for forensic analysis; platform changes, such as updates, algorithm alterations, or shutdowns, affecting data accessibility and preservation; legal requests and policies, wherein social media companies may comply with legal requests to remove certain content or user accounts, leading to the deletion or modification of relevant data; and cultural and topical shifts, where social media conversations can rapidly evolve based on current events, trends, or public sentiment, rendering older data less relevant or accurate over time. Due to this volatility, forensic analysts face challenges in collecting, storing, and analyzing data for investigative purposes. Techniques and tools for capturing and storing social media data must account for its rapid turnover and potential for modification or deletion. Additionally, forensic analysts must act swiftly to collect relevant data before it becomes inaccessible or loses its evidentiary value.

(b) Data volume and diversity: The vast amount and wide range of digital information generated and shared across social media platforms introduce a considerable level of complexity to the social media forensics process. Social media platforms host a multitude of content types, including text, images, videos, links, and more, leading to a diverse array of data formats and structures. This diversity presents challenges for

forensic analysis, as different types of content require specialized techniques for processing and interpretation. Furthermore, the sheer volume of information generated on online social networks daily is immense, making it difficult for forensic investigators to sift through and analyze relevant information efficiently. The continuous influx of data adds to the complexity, requiring forensic analysts to develop scalable methods and tools for managing and analyzing large datasets. Additionally, the global reach of social media platforms means that data can be generated in multiple languages and cultural contexts, further increasing the complexity of analysis. Therefore, effective social media forensic investigations require strategies for handling the vast volume and diverse nature of data found on these platforms, ensuring that relevant information is identified, extracted, and interpreted accurately [5].

(c) Attribution and anonymity: There are major difficulties related to identifying the creators or originators of content and distinguishing between genuine users and those hiding behind pseudonyms or false identities. Attribution involves tracing digital content back to its source or author [45], which can be challenging due to the ease of creating anonymous accounts and the potential for content to be shared and reposted across multiple platforms. Social media platforms often allow users to create accounts with minimal verification, enabling individuals to hide their true identities or impersonate others [46]. This anonymity complicates forensic investigations by obscuring the trail of digital evidence and making it difficult to establish the credibility and authenticity of information. Moreover, malicious actors may deliberately manipulate or distort information to mislead investigators or incite conflict, further complicating the task of attribution. Forensic analysts must employ advanced techniques, such as digital footprint analysis, linguistic analysis, and network analysis, to attribute digital content to its source and differentiate between legitimate users and impostors. Additionally, legal and ethical considerations surrounding user privacy and data protection must be carefully navigated when attempting to uncover the identities of individuals behind anonymous accounts. Therefore, addressing the challenges of attribution and anonymity in social media requires a mix of domain expertise, investigative rigor, and adherence to ethical standards to guarantee the accurate and responsible use of digital evidence in forensic contexts.

(d) Privacy concerns: Privacy concerns in social media forensics encompass the ethical and legal dilemmas resulting from the investigation and analysis of digital evidence gathered from social media platforms. As forensic analysts extract and scrutinize data from social media accounts, they confront the challenge of balancing the imperative to uncover truth with the imperative to protect individual privacy rights. The very nature of social media forensics, which involves accessing and examining personal information shared by users, raises concerns about the invasion of privacy and potential misuse of sensitive data. Individuals may feel uneasy knowing that their online activities are subject to scrutiny and may fear the implications of their digital footprint being used in investigations [47]. Moreover, the handling of social media data by forensic experts must adhere to strict ethical guidelines and legal regulations to safeguard the privacy of individuals and ensure the sanctity of the investigative process. Concerns also extend to the potential for data breaches or leaks during the forensic analysis, which could expose personal information to unauthorized parties and lead to further privacy violations. Thus, social media forensics practitioners face the challenge of navigating these privacy concerns while fulfilling their investigative duties. They must employ robust data protection measures, obtain appropriate legal permissions, and prioritize the anonymization of personal information whenever possible. Additionally, fostering transparency and accountability in social media forensic practices is essential for building trust with stakeholders and mitigating privacy-related apprehensions.

3.2. Traditional Social Media Forensic Techniques

Traditional social media forensic techniques encompass established methods for investigating and analyzing digital evidence sourced from social media platforms. These techniques have evolved in response to the growing significance of social media in various spheres of life. One cornerstone of traditional techniques is data collection, which involves systematically gathering digital artifacts such as user profiles, posts, comments, messages, and metadata. While traditional methods may include manual scraping or specialized software tools, they are often limited in scalability, efficiency, and coverage due to the sheer volume and dynamic form of content on social media.

Once data are collected, traditional techniques rely on manual examination and interpretation by forensic analysts. Analytical methods such as keyword analysis, sentiment analysis, and network analysis are applied to extract insights and establish connections relevant to the investigation. The core traditional techniques often utilized include:

(a) Data acquisition: Traditional methods like keyword searches and targeted data extraction from user profiles and posts serve as the bedrock for acquiring relevant evidence.
(b) Metadata analysis: Forensic investigators examines embedded timestamps, location data, and other metadata associated with the acquired social media content to better understand the context and origin of the information.
(c) Hashing and digital forensics tools: Ensuring data integrity and chain of custody is important. Forensics analysts utilize hashing algorithms and specialized software designed for traditional digital forensics.
(d) Network analysis: Network analysis is used to identify connections, groups, and interactions between users, particularly through friend lists and communication logs. This can reveal patterns and potential criminal networks, building upon established network forensics techniques.
(e) Content analysis: Traditional text analysis techniques, including keyword searches, sentiment analysis, and topic modeling, offer a starting point for understanding the content of social media posts, images, and videos.

However, this kind of manual examination is tedious and error-prone, making it challenging to keep pace with the rapid flow of social media data. Moreover, these traditional techniques face limitations in preserving data integrity and ensuring verifiability throughout the investigative process. The dynamic nature of social media content, coupled with the potential for data manipulation or deletion by users, poses challenges in maintaining the integrity of digital evidence. Chain of custody procedures and data storage methods may not always suffice to address these challenges effectively. To overcome these limitations, the application of artificial intelligence (AI) and its associated concepts emerges as a promising frontier in social media forensics. AI has the capability to automate and improve different aspects of the forensic procedure. AI-powered algorithms can process huge amounts of social media data at scale, identify patterns, detect anomalies, and extract meaningful insights more efficiently than manual methods.

The integration of AI represents a transformative shift in the field, offering the potential to address the inherent limitations of traditional techniques and unlock new capabilities for social media forensic analysis in the digital age. By leveraging AI, forensic analysts can improve the precision and effectiveness of social media forensic investigations, ultimately enhancing their ability to uncover truth and deliver justice. In subsequent sections, we delve deeply into the contemporary utilization of some of these AI concepts within the realm of social media forensics, examining their practical applications and reviewing pertinent literature. But first, we explore the application of Open-source intelligence (OSINT) tools and techniques in social media forensics.

3.3. OSINT in Social Media Forensics

OSINT, open-source intelligence, is the collection and analysis of data obtained from public or open sources like social networking sites, internet forums, or blogs. The process involves employing open-source tools to gather and assess these open information, cen-

tering on publicly accessible messages, updates, dialogues, social engagement, metadata, and diverse multimedia elements like images and videos. In the context of social media forensics, OSINT can be applied in investigations to collect data such as user profiles, posts, comments, photos, and videos. Forensic analysts can then leverage these data to identify potential suspects, track their activities, establish connections between individuals or groups, and gather evidence for legal proceedings. The benefits of OSINT in social media forensics investigations are significant—a major one is the fact that it utilizes publicly available information, saving the cost and time of tedious data acquisition processes.

OSINT techniques that may be applied in social media forensics include but are not limited to the following:

(a) Profile exploration: Examining user profiles, including bios, posts, comments, and follower lists can reveal details about a person's activities, interests, and connections.
(b) Keyword/hashtag searching: Utilizing relevant keywords and hashtags can lead investigators to discussions, photos, and videos related to the investigation.
(c) Geolocation analysis: Many social media posts contain embedded geolocation data, providing valuable insights into physical locations associated with an event or user.
(d) Social network analysis: Mapping connections between accounts and analyzing interactions within online communities can reveal patterns and identify potential collaborators or associates.

These techniques help forensic investigators systematically gather and analyze digital footprints left by individuals online; thus, helping them reconstruct events, uncover motives, and build a comprehensive understanding of suspects' online behavior. These OSINT techniques are also integrated into more advanced methodologies for social media forensics, including those based on artificial intelligence (AI), which we explore in further detail later on.

4. NLP in Social Media Forensics

NLP (natural language processing) is an area of artificial intelligence (AI) that aims to empower computers with the capability to comprehend, interpret, and reproduce human language in a way that is both meaningful and useful. NLP employs various techniques including text cleaning and preprocessing, which involves preparing textual data by eliminating noise, rectifying errors, and tokenizing words into meaningful units. Part-of-speech tagging is utilized to identify the grammatical role of each word, such as nouns, verbs, or adjectives, aiding in the comprehension of sentence structure and meaning. Named-entity recognition extracts key entities like people, organizations, and locations mentioned in the text, connecting them to real-world knowledge bases. Sentiment analysis is employed to categorize the emotional tone of text to assess user opinions, attitudes, and potential threats. Lastly, topic modeling uncovers recurring themes and topics within extensive datasets, unveiling hidden patterns and trends.

In the vast ocean of social media data, textual content reigns supreme, encompassing posts, comments, messages, and countless other forms of user-generated expression. Analyzing this textual data effectively is crucial for social media forensics, and that is where NLP emerges as a powerful ally. A range of studies have demonstrated the potential of NLP in social media forensics. Refs. [48,49] both highlight the use of NLP in social media forensic analysis, with Ref. [48] emphasizing the role of NLP in data collection and analysis, and Ref. [49] presenting a platform that outperforms other approaches in terms of precision and F1-score. Refs. [50,51] focus on specific applications of NLP in social media forensics, with Ref. [50] using a naïve Bayes classifier to identify potential lawbreakers based on their social media posts, and Ref. [51] using NLP to detect denial-of-service attacks and analyze public reactions to network outages. These studies collectively underscore the potential of NLP in enhancing the effectiveness of social media forensics.

In this section, we will explore how NLP has been applied in social media forensics to mitigate radicalization, cyberbullying, and the proliferation of fake profiles. Table 2

summarizes our review by providing an overview of the research papers explored and the methodologies they utilized.

Table 2. NLP in social media forensics.

Application	Research Studies	Methodologies Used
Radicalization Detection	[52–56]	Text preprocessing; feature extraction; word embedding; ML classification algorithms
Cyberbullying Detection	[57–59]	Word similarity and text detection; feature extraction; word embeddings; ML classification algorithms
Fake Profile Detection	[60–63]	Text modeling with BoW; dimensionality reduction; feature extraction; ML classification algorithms

4.1. Radicalization Detection

The ubiquitousness and accessibility of online social networks have completely transformed human communication, enabling the creation of virtual international communities. Individuals can now connect with people from across the globe instantly, transcending geographical boundaries and facilitating unprecedented levels of interaction. While the benefits of this are undeniable, it has also become a significant tool for violent extremists and radical belief supporters to propagate their ideologies and recruit people to their cause [64,65]. Due to the rapid spread and extensive reach of hateful content online, radical and extremist posts often propagate widely compared to other types of content [66]. This rise in online radicalization and extremism poses a significant threat to global security. To combat this rise, forensic analysts utilize natural language processing (NLP) techniques as a potential tool for identifying radicalization posts on social networks.

In leveraging natural language processing (NLP) techniques for radicalization detection, researchers explore a range of applications aimed at analyzing extremist content and language patterns. These include text preprocessing and feature engineering, which involves cleaning and preparing text data while identifying relevant features such as word frequency and sentiment, as evidenced by Refs. [52–54]. These studies use the term frequency–inverse document frequency (TF-IDF) statistical method to identify and extract the relevant features in the dataset. Additionally, word embeddings (particularly word2vec) are employed to transform words into numerical vectors, capturing semantic relationships and aiding in the identification of similar and potentially extremist content, as observed in Refs. [65–67]. Machine learning algorithms are then utilized to categorize textual data as potentially extremist based on the extracted features [55,56,67]. These research endeavors highlight the potential of NLP techniques, with Ref. [55] uncovering radical properties through language modeling and psychological profiling, Ref. [56] implementing a radicalization score and machine learning algorithms to identify potentially radicalized individuals on social media platforms, and Ref. [67] exploring right-wing extremism on social media platforms using TF-IDF and artificial neural networks for content classification.

4.2. Cyberbullying Detection

In recent years, cyberbullying, which refers to bullying conducted online or via mobile devices, has seen an increase in prevalence [68]. Cyberbullying refers to the use of electronic media like social media platforms and internet forums to intimidate, harass, or harm vulnerable individuals or groups [69]. Although there are differences among different types, cyberbullying behaviors often mimic the emotional, interpersonal, and subtle aspects of offline bullying. These behaviors can include spreading rumors, harassment, threats, and exclusion [70]. Cyberbullies commonly maintain anonymity through temporary email and fake messaging profiles, anonymizers, and pseudonyms on social networking sites, chat rooms, and forums. However, research suggests that most cyberbullying victims are

aware of who their harassers are, or at least perceive them to be members of their social circle [71,72].

The issue of identifying and addressing abusive content, as well as its various subcategories such as hate speech, toxicity, and cyberbullying, has become a focal point in the NLP research field. This topic has attracted significant attention and interest. Prioritizing ethical principles like privacy, transparency, security and fairness, Kiritchenko et al. (2020) [73] surveyed a substantial body of NLP research on automatic abuse detection, focusing on the ensuing ethical challenges. Machine learning and natural language processing are two of the most significant cutting-edge strategies that lower the incidence of cyberbullying in today's increasingly digital society. The study by Manogaran et al. (2021) [57] introduces a new detection algorithm based on word similarity and quick text detection for natural language processing using deep structured learning. Automatic detection utilizing NLP is a necessary first step that helps prevent cyberbullying, thus the significance of Ref. [57]'s detection algorithm. Utilizing NLP methods and a range of machine learning classifiers, Ahmed et al. (2021) [58] develop a model for detecting cyberbullying in Bangla and Romanized Bangla texts using comments from YouTube videos. The study used TF-IDF to extract features from the comments dataset before feeding the extracted features into a machine learning model. Elsafoury et al. (2022) [59] conducted a comparative analysis of word embeddings from social media and non-social media data on two social NLP tasks: measuring social bias and detecting cyberbullying. Their findings indicate that social-media-based word embeddings, as opposed to non-social-media-based embeddings, produce better results.

4.3. Fake Profile Detection

The ubiquitous expansion and adoption of social media platforms has spawned a parallel realm of digital deception. Within this realm, fake profiles—comprising automated social bots, also known as Sybils, impersonators, and fictitious identities—infiltrate user networks, often with malicious intent [74]. The estimated number of fake profiles on the major social networks is illustrated in Figure 4. According to Ref. [75], fake profiles are extensively utilized in perpetrating sophisticated financial scams, such as pig butchering scams, which have collectively defrauded hundreds of thousands of Americans, amounting to approximately USD 429 million [76]. The anonymity and ease of creating such profiles have significantly contributed to their proliferation. Typically, a user on OSNs is identified by a profile containing a picture, name, and possibly additional details like address and birth date. However, Ref. [45] contends that these sites often lack stringent verification mechanisms to confirm whether the individual referenced in the profile genuinely created and manages it. Studies highlight the alarming prevalence of fake profiles, with estimates suggesting millions operate across various social media platforms. These deceptive entities serve as conduits for disseminating misinformation, perpetrating financial scams, engaging in identity theft, perpetrating sexual harassment, and orchestrating targeted manipulation, thereby posing a major threat to the cyber-security and welfare of users [77].

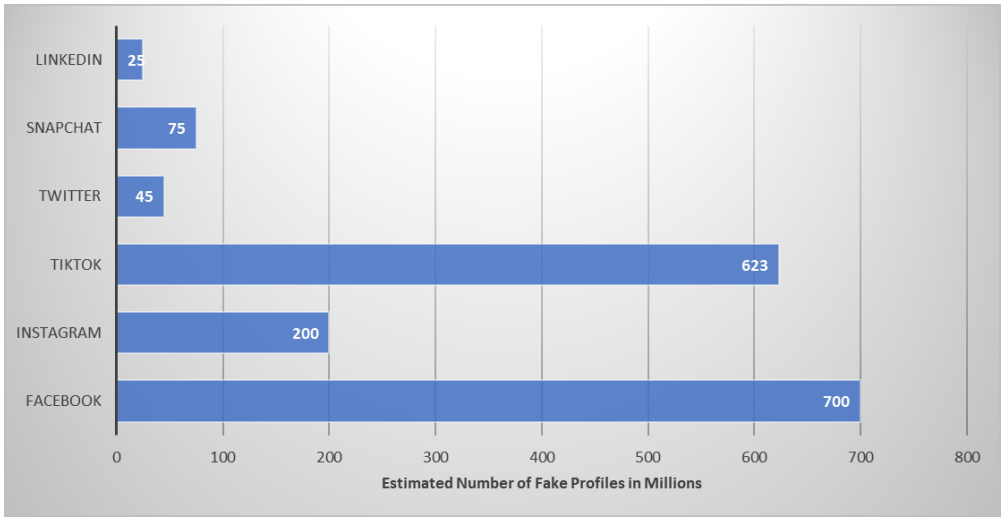

Figure 4. Fake profiles on social networks [78].

Due to this pervasiveness of fake profiles on these social media platforms, their detection is a subject of interest, with research focusing on leveraging machine learning and NLP techniques to improve accuracy rates in identifying fraudulent accounts. Several studies we reviewed [60–63] propose a combined approach involving various NLP preprocessing steps followed by ML algorithms to enhance performance in detecting fake profiles. The typical workflow encompasses three key stages: NLP preprocessing, which involves cleaning and transforming textual data through tasks like tokenization, stop word removal, stemming and lemmatization; principal component analysis (PCA) for dimensionality reduction and feature extraction (using TF-IDF) from processed data; and finally, the application of the ML algorithms to identify patterns and predict whether a given account is real or fake. Ref. [62] focuses on system architecture, proposing two distinct approaches. The first architecture employs NLP and network identifiers to identify account details. Accounts flagged by multiple users trigger a security verification process. In the second architecture, support vector machine (SVM) with the bag-of-words (BoW) concept is utilized to identify harmful words within accounts, forming a dataset for training and testing. The system calculates the frequency of harmful words in individual accounts and issues warnings for authentication if necessary, emphasizing the need for authentication before data publication. By leveraging NLP techniques, Ref. [63] built a model that could detect the authenticity of a profile from any online social network with an accuracy of up to 95%.

5. GNNs in Social Media Forensics

Graph neural networks (GNNs) are increasingly being utilized in various domains, including social media forensics, due to their ability to effectively model relational data. GNNs operate on graph structures [79], which are particularly well suited for representing complex relationships inherent in social media networks. Unlike traditional neural networks that process data with fixed dimensions, GNNs can handle data with variable structures by capturing dependencies and interactions among entities in the graph [80]. At the core of GNNs is the concept of message passing [81], where information is exchanged between nodes in the graph iteratively to update their representations. This enables GNNs to aggregate information from neighboring nodes and learn meaningful representations that encode the structural information within the graph. As a result, GNNs excel at node classification, link prediction, and graph classification, which are integral to social media forensics.

In the context of social media forensics, GNNs offer several advantages. They can effectively capture the intricate relationships between users, posts, comments, and other entities in social media networks, enabling more nuanced analysis and inference. By leveraging the rich relational information present in social media graphs, GNNs can

uncover patterns of behavior, detect anomalies, and identify malicious activities such as cyberbullying, fake account propagation, and extremist recruitment. GNNs are inherently scalable and adaptable to different types of social media networks, ranging from small-scale communities to large-scale platforms with millions of users. Their ability to generalize across diverse graphs makes them suitable for addressing various challenges in social media forensics, including data sparsity, noise, and evolving network structures.

In subsequent subsections, we discuss the applications of GNNs in social media forensics. We highlight key research trends and recent advancements in the integration of GNNs into forensic analysis workflows, paving the way for more effective and efficient investigations in the realm of social media.

5.1. Fauxtography

Fauxtography, the dissemination of manipulated or misleading images across social media platforms, presents a significant challenge in social media forensics. This is because manipulated images negatively impact the validity and integrity of the data collected by forensic analysts in the course of their investigations. Ref. [82] described fauxtography as picture(s) and text associated with a social media post that collectively portray events in a questionable or entirely fabricated manner.

GNNs provides a promising approach for detecting fauxtography due to their ability to capture intricate patterns within image datasets. At the heart of GNN-based fauxtography detection lies the representation of images and their associated metadata as a graph structure. In this framework, each image is represented as a node in the graph, while relationships between images, such as similarity or co-occurrence, are represented as edges. By leveraging the inherent connectivity between images, GNNs can effectively learn patterns that reveals the fundamental structure of the image dataset. One key advantage of GNNs in fauxtography detection is their ability to incorporate both visual features extracted from images and contextual information derived from metadata, such as timestamps, geotags, and user interactions. This holistic representation enables GNNs to capture subtle cues indicative of image manipulation or misinformation, such as inconsistencies in timestamps or anomalous patterns of user engagement.

One of the pioneering studies to utilize GNNs in fauxtography was the paper by Zhang et al. (2018) [83]. They introduced a novel approach, FauxBuster, for detecting fauxtography on social media by analyzing user comments rather than image content. This content-free method was shown to be effective and efficient in tracking down misleading information conveyed through images and texts on platforms like Reddit and Twitter. FauxBuster outperformed existing image forgery detection methods, achieving a 25.6% higher F1-score and an 86.1% detection accuracy. The authors utilized graph neural networks (GNNs) for network representation, learning through a stacked autoencoding technique. By extracting feature vectors from random walks and deriving the comment network's signature using deep autoencoders, the GNN was employed to learn abstract features from high-dimensional data in an unsupervised manner. This allowed for the reduction of complexity in input data and the mapping of input vectors into a latent subspace for improved analysis and understanding of the comment network's structure and characteristics. Ref. [84] developed a framework called FauxWard, which is a graph convolutional neural network approach designed to detect fauxtography on social media based on social media comments. The methodology employed delves into the intricate data gathered from a network of user comments to accurately detect fauxtography posts. Because FauxWard does not examine the text or image content of the post, it can withstand sophisticated fauxtography uploaders who purposefully manipulate posts to appear misleading by manipulating the text or image content. By using graph convolutional layers with activation functions and message propagation functions to aggregate node information, the authors effectively encode user comment networks and represent the problem as a graph classification problem using graph convolutional neural networks (GCNNs).

5.2. Criminal Activity Detection

GNNs serve as a potent tool in the realm of social media forensics for detecting wide varieties of criminal activity, including drug trafficking and troll behavior. The interconnected nature of social media networks lends itself well to representation as graphs, where nodes represent users or entities, and edges capture relationships or interactions between them. Forensic researchers are beginning to use GNNs to leverage this rich structural information to analyze patterns of behavior and identify anomalous or illicit activities within social media ecosystems.

Qian et al. (2021) [85] propose a framework called MetaHG to detect drug traffickers on social networks, specifically Instagram. Their methodology involves building a heterogeneous graph (HG) to represent the intricate network of drug trafficking on social media. This is followed by the use of a relation-based graph convolutional neural network (GCNN) to learn node representations over the built HG, thus improving the graph's node representation learning. They utilize relation-based graph convolutional neural networks (R-GCNs) to fuse relational information among entities and content features, obtain initial node embeddings on an HG, and further refine node representations via graph structure refinement (GSR) to comprehensively characterize drug trafficking on social media. Additionally, they incorporate a self-supervised learning module for node representation refinement and a knowledge distillation module for model optimization in exploiting unlabeled data for improved performance. Ref. [86] pioneers implementing deep learning algorithms in graph convolutional networks to map out patterns associated with trolls or harmful material shared on social media websites. The research introduces the use of GCNNs to examine and analyze the intricate patterns within social media data. Their GCNN-based framework reported high accuracy scores, with -0.92 testing accuracy and 0.88 inference accuracy. This is a marked improvement on similar research models that relied on LSTM architectures.

6. GANs in Social Media Forensics

The critical need for advanced forensic techniques to effectively combat digital deception has been underlined in the past few years by the rapid spread of misinformation, fake news, and manipulated content on social media platforms. Generative adversarial networks (GANs) have emerged as a promising AI technology in the field of social media forensics, offering innovative solutions to address the complexities of detecting and combating deceptive practices. It is important to state that GANs are often employed to create these deceptive, hyper-realistic forgeries. In this section, however, we will examine how they can also be employed to detect and combat the proliferation of these forgeries on online social networks.

GANs are a framework for artificial intelligence models composed of two neural networks, a discriminator and a generator, that are concurrently trained via a competitive process [87]. Figure 5 succinctly depicts their basic working principle. The generator network (usually a convolutional neural network) creates new data samples, such as images, based on random noise or other input data. In contrast, the discriminator network (usually a deconvolutional neural network) distinguishes between real and fake samples [88]. Through this adversarial training process, GANs learn to generate increasingly realistic samples, pushing the boundaries of what can be achieved regarding creativity and realism, with applications ranging from image generation and style transfer to data augmentation and synthetic data generation for training machine learning models.

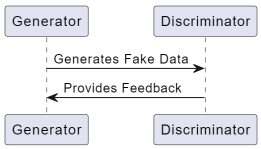

Figure 5. Basic working principle of GANs.

This section explores the integration of GANs within social media forensics, focusing on their application in detecting deepfakes. By harnessing the power of adversarial learning, GANs enable forensic analysts to identify and mitigate a wide array of digital manipulations, ranging from fake images and videos to fabricated text and audio content. Moreover, GANs offer a novel approach to generating realistic synthetic data, facilitating the creation of diverse datasets crucial for training robust forensic models and improving detection accuracy.

Deepfake Detection

In social media forensics, the emergence of deepfake technology poses a significant challenge to the integrity of digital content, amplifying the spread of misinformation and undermining trust in online platforms. Addressing this pressing issue requires innovative approaches, and one such solution lies in utilizing generative adversarial networks (GANs) for deepfake detection.

Deepfake refers to manipulated media, typically videos, that use advanced artificial intelligence techniques, such as deep learning algorithms and GANs, to convincingly depict events or speeches that never occurred or were never said by the individuals depicted [89,90]. These manipulated videos and pictures can be used for different purposes, including creating fake events, altering speeches, and generating non-consensual explicit content [90]. Figure 6 provides a sample of two deepfake images created to duplicate the likeness of two popular American movie stars. Deepfake technology has raised concerns about its potential abuse, driving researchers and forensic analysts to explore novel detection methods to curb the dissemination of such manipulated media.

Figure 6. Original images and their deepfake versions [91].

GANs have long played a pivotal role in the proliferation of deepfake technology by enabling the creation of highly realistic synthetic media that can deceive unsuspecting viewers. Various deepfake creation tools (FaceSwap, StyleGAN, DeepFaceLab, DiscoFaceGAN, etc.) utilize GAN-based architectures to generate fake media. However, as the threat of deepfakes continues to escalate, researchers and forensic analysts are shifting their focus towards leveraging GANs for detection rather than solely creation. By re-purposing

GANs and exploring innovative detection strategies, researchers aim to enhance the resilience of digital media ecosystems against the threat of deceptive manipulation, thereby safeguarding the integrity of online content and restoring trust in visual information.

In Ref. [92], the researchers employed GAN for deepfake detection by utilizing a customized deep convolution GAN architecture tailored specifically for this purpose and leveraging the CelebA dataset [93]. The model is constructed based on the principles and methodologies outlined in a prior [94] study. The paper compares their suggested GAN model against existing GAN models, assessing parameters such as inception score (IS) and Fréchet inception distance (FID) and recording optimal scores of 1.074 and 49.3, respectively. Yang et al., in their study [95] propose a defense mechanism against GAN-based deepfake attacks by using transformation-conscious adversarial faces. This method creates new adversarially perturbed faces with random image alterations during generation. When adversarial faces are used to train a deepfake model, the quality of the resulting synthesized face degrades significantly, rendering it more blatantly fake and susceptible to detection. The study proactively tries to prevent the creation of hyper-realistic high-quality fake images or videos, particularly targeting GAN-based deepfake attacks. In Ref. [96], Nadimpalli and Rattani introduce an innovative preemptive method for detecting deepfakes by utilizing GAN-based visible watermarking, aiming to thwart the misuse of deepfakes for nefarious purposes. By incorporating a distinct watermark into fabricated images during generation, the technique seeks to facilitate effortless identification of deepfakes by human observers and advanced detection systems. This proactive strategy is a robust defense mechanism against deepfakes, overcoming the constraints of current passive detection methods. Ref. [97] proposes a novel deepfake detection technique, which they brand the CTF method. The study suggests that by analyzing discrete cosine transform (DCT) coefficient statistics, it is possible to distinguish GAN-based deepfakes from original content using the GAN specific frequency band (GSF). GSF exhibits several noteworthy properties, including aiding in understanding the deepfake generation process, particularly for forensic applications.

7. Challenges and Future Directions

In this section, we will highlight the current challenges plaguing the application of artificial intelligence techniques in social media forensics. We will also discuss potential directions for future researchers and forensics practitioners.

7.1. Key Challenges

(a) Data availability and privacy: Balancing the need for comprehensive data for effective AI model training and analysis with the paramount importance of user privacy remains a significant hurdle. Collaborations between researchers, law enforcement agencies, and social media platforms are crucial to establish ethical and legal frameworks for data access while upholding user privacy rights.

(b) Explainability and interpretability: The "black box" nature of many AI models, particularly complex algorithms like deep learning architectures, raises concerns about their decision-making processes. Developing interpretable AI techniques is vital for building trust and ensuring ethical application in forensic investigations. This requires advancements in model design and the integration of Explainable AI (XAI) methodologies to offer insights into the process by which AI models derive their conclusions.

(c) Bias and fairness: AI models have the capacity to adopt and magnify biases inherent in their training data, possibly resulting in unjust or prejudiced results. Mitigating bias requires comprehensive approaches, including:

   i. Employing diverse datasets: Utilizing data that reflects the true diversity of online communities is crucial to avoid perpetuating existing biases.
   ii. Developing fair evaluation metrics: Establishing evaluation metrics that not only assess accuracy but also identify and address potential biases within the model's predictions.

iii. Careful model design: Implementing techniques like fairness-aware model architectures and training procedures can help mitigate bias from the outset.

(d) Evolving technologies and user behavior: The rapid pace of technological advancements and user behavior changes necessitate continuous adaptation and refinement of AI models. Continuously updating training data, developing generalizable models, and monitoring their performance in real-world scenarios are essential to ensure effectiveness and avoid model drift.

7.2. Future Directions

The integration of various AI techniques, including NLP and GNNs, into social media forensics holds immense promise for enhancing investigative capabilities. To better deliver on this promise, we highlight potential areas where research efforts should be concentrated.

(a) Interdisciplinary collaboration: Future research in AI-driven social media forensics should foster interdisciplinary collaboration between computer scientists, social scientists, legal experts, and ethicists. Collaborative efforts can facilitate a more holistic understanding of the complex socio-technical challenges involved in forensic investigations.

(b) Explainable AI (XAI): Improving the explainability and interpretability of AI models is crucial for fostering trust and transparency in forensic decision-making processes. Future research should prioritize the development of XAI techniques capable of providing human-understandable explanations for AI-driven forensic analyses.

(c) Continuous learning and adaptation: AI systems in social media forensics should be designed to learn continuously from new data and adapt to evolving threats and challenges. Incorporating mechanisms for online learning and real-time feedback can enhance the agility and effectiveness of forensic analyses in dynamic social media environments.

(d) Privacy-preserving techniques: Advancing privacy-preserving AI techniques is paramount for safeguarding user privacy while enabling effective forensic analyses. Future research should explore innovative approaches for conducting forensic investigations while minimizing the disclosure of sensitive user information.

(e) Ethical guidelines and standards: It is imperative to establish unambiguous ethical guidelines and standards to govern the conscientious application of artificial intelligence in the field of social media forensics. Future efforts should focus on developing ethical frameworks and regulatory mechanisms to ensure the ethical and responsible deployment of AI technologies in forensic investigations.

7.3. Limitations of the Scope

This survey, titled "Artificial Intelligence in Social Media Forensics: A Comprehensive Survey and Analysis", strives to provide a thorough examination of the transformative potential of AI in social media forensics. However, the dynamic nature of AI research and the ever-expanding possibilities within social media forensics necessitate acknowledging the inherent limitations of achieving complete comprehensiveness in a single work. It is important to recognize that this survey cannot exhaustively cover every facet of AI in social media forensics. The field is constantly witnessing advancements in AI techniques and their applications. Additionally, the sheer volume of ongoing research makes it challenging to capture every single contribution. We have chosen to focus on the most prominent and well-established areas of AI application: (1) natural language processing (NLP) for fake profile detection, cyberbullying analysis, and radicalization identification; (2) graph neural networks (GNNs) for uncovering hidden connections in social networks, aiding in investigations of online fraud and criminal activity; and (3) generative adversarial networks (GANs) for deepfake detection, a growing challenge in the digital space. By focusing on these key areas, we aim to provide a solid foundation for understanding the current landscape and future directions of AI in social media forensics. However, we acknowledge the existence of other promising AI techniques and applications that deserve further exploration. We hope this survey serves as a stepping stone for further exploration of the ever-expanding possibilities of AI in social media forensics.

8. Conclusions

As the world grows increasingly digital, and the bulk of our interactions move online—on social media platforms and internet forums—it becomes imperative for us to create accountability and security mechanisms for this digital brave new world. Social media forensics is a budding research field seeking to aid the investigation and punishment of harmful and antisocial behavior on these social networks. Thus, ensuring that even as we move online, our values, laws, and processes move online with us.

This paper explores how artificial intelligence concepts are utilized in the social media forensics field. We start off by discussing digital forensics, analyzing its subdomains and highlighting recent research works and notable advancements in each field. We identify NLP, GNNs, and GANs as three leading AI techniques utilized in the social media forensics field. We review the existing literature, methodologies, and tools in the field to explore how these methods are used to combat common cybercrimes encountered by forensic investigators. This paper highlights the potential of AI methodologies in social media forensics investigations. The literature surveyed demonstrates how deep learning and NLP techniques are applied to finding, collecting, and analyzing digital evidence on social networking websites.

This paper, in addition to acknowledging the challenges involved and ethical considerations tied to integrating AI into forensics procedures, underscores the importance of researchers and investigators adhering strictly to data collection laws and privacy guidelines. It thereby accentuates a critical need for AI models to prioritize bias mitigation while concurrently establishing trust with individuals. We recommend that future research works not only address these aforementioned challenges but also delve deeper into their implications; this is an imperative course of action if we are truly committed towards advancing responsibly within this burgeoning field. We also advocate for fostering interdisciplinary collaboration in the field of social media forensics, which inherently embodies a multidisciplinary nature: it combines computer science, law, and social governance. This paper contributes to our collective objective of engendering a safer, more secure digital environment for everyone.

Author Contributions: Conceptualization, B.G.B. and Q.L.; Methodology, B.G.B. and Q.L.; Validation, B.G.B.; Formal analysis, B.G.B.; Investigation, B.G.B.; Writing—original draft, B.G.B.; Writing—review & editing, Q.L.; Supervision, Q.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the manuscript.

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

References

1. Dean, B. *Social Network Usage & Growth Statistics: How Many People Use Social Media in 2024?* Backlinko: Cheyenne, WY, USA, 2024. Available online: https://backlinko.com/social-media-users (accessed on 19 April 2024).
2. The Importance and Challenges of Social Media in Digital Investigations. Available online: https://www.controlrisks.com/our-thinking/insights/the-importance-and-challenges-of-social-media-in-digital-investigations?utm_referrer=https://www.google.com (accessed on 19 April 2024).
3. Dwivedi, Y.K.; Kelly, G.; Janssen, M.; Rana, N.P.; Slade, E.L.; Clement, M. Social Media: The Good, the Bad, and the Ugly. *Inf. Syst. Front.* **2018**, *20*, 419–423. [CrossRef]
4. Morgan, S. Cybercrime to Cost the World 8 Trillion Annually in 2023. *Cybercrime Magazine*, 17 October 2022. Available online: https://cybersecurityventures.com/cybercrime-to-cost-the-world-8-trillion-annually-in-2023/ (accessed on 19 April 2024).
5. *Digital Forensics and Social Media: Ethics, Challenges and Opportunities*; Birkbeck, University of London: London, UK, 2019. Available online: https://www.bbk.ac.uk/news/digital-forensics-and-social-media-ethics-challenges-and-opportunities/ (accessed on 19 April 2024).
6. Kent, K.; Chevalier, S.; Grance, T.; Dang, H. *Special Publication 800-86 Guide to Integrating Forensic Techniques into Incident Response Recommendations of the National Institute of Standards and Technology*; The National Institute of Standards and Technology: Gaithersburg, MD, USA, 2006. Available online: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-86.pdf (accessed on 3 February 2024).

7. Sharma, B.K.; Joseph, M.A.; Jacob, B.; Miranda, B. Emerging trends in digital forensics and cybersecurity—An overview. In Proceedings of the 2019 Sixth HCT Information Technology Trends (ITT), Ras Al Khaimah, United Arab Emirates, 20–21 November 2019; pp. 309–313.
8. Dumchykov, M. The Processes of Digitization and Forensics: A Retrospective Analysis. *Crim. Forensics* **2020**, *65*, 100–108. [CrossRef]
9. Ivanov, V.Y. On theoretical aspects of using the concept of digital footprint in forensics. *Leg. Stud.* **2020**, 75–80. [CrossRef]
10. Sachowski, J. *Implementing Digital Forensic Readiness: From Reactive to Proactive Process*, 2nd ed.; CRC Press: Boca Raton, FL, USA, 2019. [CrossRef]
11. Roux, C.; Crispino, F.; Ribaux, O. From forensics to forensic science. *Curr. Issues Crim. Justice* **2012**, *24*, 7–24. [CrossRef]
12. Karabiyik, U. Building an Intelligent Assistant for Digital Forensics. Ph.D. Thesis, Florida State University, Tallahassee, FL, USA, 2015.
13. Kizza, J.M. Computer crime investigations—Computer forensics. In *Ethical and Social Issues in the Information Age, Texts in Computer Science*; Springer: London, UK, 2010; pp. 263–276. [CrossRef]
14. Pratama, I.P.A.E. Computer Forensic Using Photorec for Secure Data Recovery Between Storage Media: A Proof of Concept. *Int. J. Sci. Technol. Manag.* **2021**, *2*, 1189–1196. [CrossRef]
15. Khan, M.N.A.; Chatwin, C.R.; Young, R.C. A framework for post-event timeline reconstruction using neural networks. *Digit. Investig.* **2007**, *4*, 146–157. [CrossRef]
16. Huebner, E.; Bem, D.; Henskens, F.; Wallis, M. Persistent systems techniques in forensic acquisition of memory. *Digit. Investig.* **2007**, *4*, 129–137. [CrossRef]
17. Studiawan, H.; Sohel, F.; Payne, C. A survey on forensic investigation of operating system logs. *Digit. Investig.* **2019**, *29*, 1–20. [CrossRef]
18. Dalal, M.; Juneja, M. Video steganalysis to obstruct criminal activities for digital forensics: A survey. *Int. J. Electron. Secur. Digit. Forensics* **2018**, *10*, 338. [CrossRef]
19. Davidson, J.; Bergman, C.; Bartlett, E. An artificial neural network for wavelet steganalysis. In Proceedings of the Optics and Photonics 2005, San Diego, CA, USA, 31 July–4 August 2005. [CrossRef]
20. Zaharis, A.; Martini, A.; Tryfonas, T.; Ilioudis, C.; Pangalos, G. Lightweight Steganalysis Based on Image Reconstruction and Lead Digit Distribution Analysis. *Int. J. Digit. Crime Forensics* **2011**, *3*, 29–41. [CrossRef]
21. Sharma, P.; Arora, D.; Sakthivel, T. Enhanced Forensic Process for Improving Mobile Cloud Traceability in Cloud-Based Mobile Applications. *Procedia Comput. Sci.* **2020**, *167*, 907–917. [CrossRef]
22. Joseph, M.A.; Philip, S.; Miranada, B.; Deshmukh, A.; Singh, N. A theoretical workflow for the verification of embedded threats on mobile devices. In Proceedings of the 2021 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM), Dubai, United Arab Emirates, 19–21 January 2021. [CrossRef]
23. Koroniotis, N.; Moustafa, N.; Sitnikova, E. A new network forensic framework based on deep learning for Internet of Things networks: A particle deep framework. *Future Gener. Comput. Syst.* **2020**, *110*, 91–106. [CrossRef]
24. Sikos, L.F. Packet analysis for network forensics: A comprehensive survey. *Forensic Sci. Int. Digit. Investig.* **2020**, *32*, 200892. [CrossRef]
25. Khalid, Z.; Iqbal, F.; Kamoun, F.; Hussain, M.; Khan, L.A. Forensic analysis of the cisco WebEx application. In Proceedings of the 2021 5th Cyber Security in Networking Conference (CSNet), Abu Dhabi, United Arab Emirates, 12–14 October 2021. [CrossRef]
26. Lo, W.W.; Kulatilleke, G.; Sarhan, M.; Layeghy, S.; Portmann, M. XG-BoT: An Explainable Deep Graph Neural Network for Botnet Detection and Forensics. *Internet Things* **2022**, *22*, 100747. [CrossRef]
27. Khanuja, H.K.; Adane, D. Monitor and detect suspicious transactions with database forensics and Dempster-Shafer theory of evidence. *Int. J. Electron. Secur. Digit. Forensics* **2020**, *12*, 154. [CrossRef]
28. Al-Dhaqm, A.; Razak, S.; Ikuesan, R.A.; Kebande, V.R.; Hajar Othman, S. Face Validation of Database Forensic Investigation Metamodel. *Infrastructures* **2021**, *6*, 13. [CrossRef]
29. Chopade, R.M.; Pachghare, V.K. Data Tamper Detection from NoSQL Database in Forensic Environment. *J. Cyber Secur. Mobil.* **2021**, *10*, 421–450. [CrossRef]
30. Choi, H.; Lee, S.; Jeong, D. Forensic Recovery of SQL Server Database: Practical Approach. *IEEE Access* **2021**, *9*, 14564–14575. [CrossRef]
31. Zhang, C.; Yin, J. Research on security mechanism and forensics of SQLite database. In *Communications in Computer and Information Science*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 614–629. [CrossRef]
32. Rani, D.R.; Geethakumari, G. Secure data transmission and detection of anti-forensic attacks in cloud environment using MECC and DLMNN. *Comput. Commun.* **2020**, *150*, 799–810. [CrossRef]
33. Ahsan, M.M.; Wahab, A.W.B.A.; Idris, M.Y.I.B.; Khan, S.; Bachura, E.; Choo, K.K.R. CLASS: Cloud Log Assuring Soundness and Secrecy Scheme for Cloud Forensics. *IEEE Trans. Sustain. Comput.* **2021**, *6*, 184–196. [CrossRef]
34. Awuson-David, K.; Al-Hadhrami, T.; Alazab, M.; Shah, N.; Shalaginov, A. BCFL logging: An approach to acquire and preserve admissible digital forensics evidence in cloud ecosystem. *Future Gener. Comput. Syst.* **2021**, *122*, 1–13. [CrossRef]
35. U.S. Department of Homeland Security. *Computer Forensics*; U.S. Department of Homeland Security: Washington, DC, USA, 2008.

36. EC-Council. How to Handle Data Acquisition in Digital Forensics, Cybersecurity Exchange. 11 March 2022. Available online: https://www.eccouncil.org/cybersecurity-exchange/computer-forensics/data-acquisition-digital-forensics/ (accessed on 19 April 2024).
37. Pedapudi, S.M.; Vadlamani, N. Data acquisition based seizure record framework for digital forensics investigations. In Proceedings of the 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2–4 December 2021. Available online: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9676088 (accessed on 16 August 2022).
38. Christophe Grenier. Photorec. Available online: http://www.cgsecurity.org/wiki/photorec (accessed on 19 April 2024).
39. Mohammad, R.M. A neural network based digital forensics classification. In Proceedings of the 2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA), Aqaba, Jordan, 28 October–1 November 2018. [CrossRef]
40. Garfinkel, S.L. Digital forensics research: The next 10 years. *Digit. Investig.* **2010**, *7*, S64–S73. [CrossRef]
41. Gallagher, S. *Steganography: How Al-Qaeda Hid Secret Documents in a Porn Video*; Ars Technica: New York, NY, USA, 2012. Available online: https://arstechnica.com/business/2012/05/steganography-how-al-qaeda-hid-secret (accessed on 23 February 2024).
42. Olivier, M.S. On metadata context in database forensics. *Digit. Investig.* **2009**, *5*, 115–123. [CrossRef]
43. Al-Dhaqm, A.; Abd Razak, S.; Othman, S.H.; Ali, A.; Ghaleb, F.A.; Rosman, A.S.; Marni, N. Database Forensic Investigation Process Models: A Review. *IEEE Access* **2020**, *8*, 48477–48490. [CrossRef]
44. Karagiannis, C.; Vergidis, K. Digital Evidence and Cloud Forensics: Contemporary Legal Challenges and the Power of Disposal. *Information* **2021**, *12*, 181. [CrossRef]
45. Romanov, A.; Semenov, A.; Mazhelis, O.; Veijalainen, J. Detection of fake profiles in social media—Literature review. In Proceedings of the 13th International Conference on Web Information Systems and Technologies, Porto, Portugal, 25–27 April 2017. [CrossRef]
46. Juola, P. Authorship attribution. *Found. Trends Inf. Retr.* **2006**, *1*, 233–334. [CrossRef]
47. Naqvi, S.; Enderby, S.; Williams, I.; Asif, W.; Rajarajan, M.; Potlog, C.; Florea, M. Privacy-Preserving Social Media Forensic Analysis for Preventive Policing of Online Activities. In Proceedings of the 2019 10th IFIP International Conference on New Technologies, Mobility and Security (NTMS), Canary Islands, Spain, 24–26 June 2019. [CrossRef]
48. Shahbazi, Z.; Byun, Y.C. NLP-Based Digital Forensic Analysis for Online Social Network Based on System Security. *Int. J. Environ. Res. Public Health* **2022**, *19*, 7027. [CrossRef] [PubMed]
49. Sun, D.; Zhang, X.; Choo, K.K.R.; Hu, L.; Wang, F. NLP-based digital forensic investigation platform for online communications. *Comput. Secur.* **2021**, *104*, 102210. [CrossRef]
50. Ketcham, M.; Ganokratanaa, T.; Bansin, S. The forensic algorithm on facebook using natural language processing. In Proceedings of the 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Naples, Italy, 28 November–1 December 2016. [CrossRef]
51. Chambers, N.; Fry, B.; McMasters, J. Detecting Denial-of-Service Attacks from Social Media Text: Applying NLP to Computer Security. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018. Available online: https://aclanthology.org/N18-1147/ (accessed on 7 October 2023).
52. Mursi, K.T.; Alahmadi, M.D.; Alsubaei, F.S.; Alghamdi, A.S. Detecting Islamic Radicalism Arabic Tweets Using Natural Language Processing. *IEEE Access* **2022**, *10*, 72526–72534. [CrossRef]
53. Torregrosa, J.; Bello-Orgaz, G.; Martinez-Camara, E.; Del Ser, J.; Camacho, D. A survey on extremism analysis using natural language processing. *arXiv* **2021**, arXiv:2104.04069.
54. Ul Rehman, Z.; Abbas, S.; Khan, M.A.; Mustafa, G.; Fayyaz, H.; Hanif, M.; Saeed, M.A. Understanding the Language of ISIS: An Empirical Approach to Detect Radical Content on Twitter Using Machine Learning. *Comput. Mater. Contin.* **2021**, *66*, 1075–1090. [CrossRef]
55. Nouh, M.; Nurse, J.R.; Goldsmith, M. Understanding the radical mind: Identifying signals to detect extremist content on Twitter. In Proceedings of the 2019 IEEE International Conference on Intelligence and Security Informatics (ISI), Shenzhen, China, 1–3 July 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 98–103.
56. Oussalah, M.; Faroughian, F.; Kostakos, P. On detecting online radicalization using natural language processing. In Proceedings of the Intelligent Data Engineering and Automated Learning–IDEAL 2018: 19th International Conference, Madrid, Spain, 21–23 November 2018; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; Part II 19, pp. 21–27.
57. Manogaran, G.; Qudrat-Ullah, H.; Xin, Q. (Eds.) Special issue on deep structured learning for natural language processing. In *ACM Transactions on Asian and Low-Resource Language Information Processing*; Association for Computing Machinery: New York, NY, USA, 2021;Volume 20, pp. 1–2. [CrossRef]
58. Ahmed, M.T.; Rahman, M.; Nur, S.; Islam, A.Z.M.T.; Das, D. Natural language processing and machine learning based cyberbullying detection for Bangla and Romanized Bangla texts. *TELKOMNIKA Telecommun. Comput. Electron. Control* **2021**, *20*, 89. [CrossRef]
59. Elsafoury, F.; Wilson, S.R.; Ramzan, N. A Comparative Study on Word Embeddings and Social NLP Tasks. In Proceedings of the Tenth International Workshop on Natural Language Processing for Social Media, Seattle, WA, USA, 14–15 July 2022. Available online: https://aclanthology.org/2022.socialnlp-1.5 (accessed on 26 February 2024).

60. Latha, P.; Sumitra, V.; Sasikala, V.; Arunarasi, J.; Rajini, A.R.; Nithiya, N. Fake profile identification in social network using machine learning and NLP. In Proceedings of the 2022 International Conference on Communication, Computing and Internet of Things (IC3IoT), Chennai, India, 10–11 March 2022. [CrossRef]
61. Rao, P.S.; Gyani, J.; Narsimha, G. Fake profile identification in online social networks using machine learning and NLP. *Int. J. Appl. Eng. Res.* **2018**, *13*, 973–4562.
62. Rohit, R. Machine learning implementation for identifying fake accounts in social network. *Int. J. Pure Appl. Math.* **2018**, *118*, 4785–4797.
63. Milind, S.K.; Dhamdhere, V. Automatic Detection of Fake Profiles in Online Social Networks. In *The Techincal Writers Handbook*; Young, M., Ed.; University Science: Mill Valley, CA, USA, 1989.
64. Bowman-Grieve, L. Exploring 'stormfront': A virtual community of the radical right. *Stud. Confl. Terror.* **2009**, *32*, 989–1007. [CrossRef]
65. Sageman, M. *Leaderless Jihad: Terror Networks in the Twenty-First Century*; University of Pennsylvania Press: Philadelphia, PA, USA, 2008.
66. Mathew, B.; Dutt, R.; Goyal, P.; Mukherjee, A. Spread of hate speech in online social media. In Proceedings of the WebSci '19: 11th ACM Conference on Web Science, Boston, MA, USA, 30 June–3 July 2019; pp. 173–182.
67. Løvås, I.V. Recognizing Social Media Right-Wing Radicalization Using Text Analysis and Artificial Intelligence. Master's Thesis, NTNU, Trondheim, Norway, 2022.
68. Chen, L.; Liu, X.; Tang, H. The interactive effects of parental mediation strategies in preventingcyberbullying on social media. *Psychol. Res. Behav. Manag.* **2023**, 1009–1022. [CrossRef]
69. Smith, P.K.; Mahdavi, J.; Carvalho, M.; Fisher, S.; Russell, S.; Tippett, N. Cyberbullying: Its natureand impact in secondary school pupils. *J. Child Psychol. Psychiatry* **2008**, *49*, 376–385. [CrossRef] [PubMed]
70. Bokolo, B.G.; Liu, Q. *Combating Cyberbullying in Various Digital Media Using Machine Learning*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2023; pp. 71–97. [CrossRef]
71. Kowalski, R.M.; Limber, S.P. Electronic bullying among middle school students. *J. Adolesc. Health* **2007**, *41*, S22–S30. [CrossRef] [PubMed]
72. Hinduja, S.; Patchin, J. *Bullying Beyond the Schoolyard: Preventing and Responding to Cyberbullying*; Corwin Press: Thousand Oaks, CA, USA, 2009.
73. Kiritchenko, S.; Nejadgholi, I.; Fraser, K.C. Confronting Abusive Language Online: A Survey from the Ethical and Human Rights Perspective. *J. Artif. Intell. Res.* **2021**, *71*, 431–478. [CrossRef]
74. Ali, H.; Malik, I.; Mahmood, S.; Akif, F.; Amin, J. Sybil detection in online social networks. In Proceedings of the 2022 17th International Conference on Emerging Technologies (ICET), Swabi, Pakistan, 29–30 November 2022. [CrossRef]
75. Pig Butchering Scam: From Tinder and TikTok to WhatsApp and Telegram, How Scammers Are Stealing Millions in a Long Con, Tenable®. 2024. Available online: https://www.tenable.com/blog/pig-butchering-scam-tinder-tiktok-whatsapp-telegram-scammers-steal-millions#webinar-2/22 (accessed on 19 February 2024).
76. Abbate, P. *Federal Bureau of Investigation Internet Crime Report 2021*; Internet Crime Complaint Center: Washington, DC, USA, 2021.
77. Fire, M.; Goldschmidt, R.; Elovici, Y. Online Social Networks: Threats and Solutions. *IEEE Commun. Surv. Tutorials* **2014**, *16*, 2019–2036. [CrossRef]
78. Wolotko, D. How Many Fake Accounts Are on Social Media?—Hypetrain's Blog. Available online: https://blog.hypetrain.io/fake_accounts/ (accessed on 18 March 2024).
79. Liu, Z.; Zhou, J. Introduction to graph neural networks. In *Synthesis Lectures on Artificial Intelligence and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2020; Volume 14, pp. 1–127. [CrossRef]
80. Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Philip, S.Y. A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 4–24. [CrossRef] [PubMed]
81. Gilmer, J.; Schoenholz, S.S.; Riley, P.F.; Vinyals, O.; Dahl, G.E. Neural message passing for quantum chemistry. In Proceedings of the ICML'17: Proceedings of the 34th International Conference on Machine Learning, Sydney, NSW, Australia, 6–11 August 2017; pp. 1263–1272.
82. Cooper, S.D. A concise history of the fauxtography blogstorm in the 2006 lebanon war. *Am. Commun. J.* **2007**, *9*, 2.
83. Zhang, D.Y.; Shang, L.; Geng, B.; Lai, S.; Li, K.; Zhu, H.; Amin, M.T.; Wang, D. Fauxbuster: A content-free fauxtography detector using social media comments. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 891–900.
84. Shang, L.; Zhang, Y.; Zhang, D.; Wang, D. Fauxward: A graph neural network approach to fauxtography detection using social media comments. *Soc. Netw. Anal. Min.* **2020**, *10*, 76. [CrossRef]
85. Qian, Y.; Zhang, Y.; Ye, Y.; Zhang, C. Distilling meta knowledge on heterogeneous graph for illicit drug trafficker detection on social media. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 26911–26923.
86. Asif, M.; Al-Razgan, M.; Ali, Y.A.; Yunrong, L. Graph convolution networks for social media trolls detection use deep feature extraction. *J. Cloud Comput.* **2024**, *13*, 33. [CrossRef]
87. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Montreal Canada, 8–13 December 2014; pp. 2672–2680.

88. What is a Generative Adversarial Network (GAN)? Definition from TechTarget. Enterprise AI. Available online: https://www.techtarget.com/searchenterpriseai/definition/generative-adversarial-network-GAN# (accessed on 8 March 2024).
89. Wikipedia Contributors. "Deepfake". 2019. Available online: https://en.wikipedia.org/wiki/Deepfake (accessed on 19 April 2024).
90. Sample, I. What Are Deepfakes—And How Can You Spot Them? *The Guardian*, 13 January 2020. Available online: https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them (accessed on 19 April 2024).
91. File: Deepfake Metahuman.png—Wikimedia Commons. 2024. Available online: https://commons.m.wikimedia.org/wiki/File:Deepfake_Metahuman.png (accessed on 18 March 2024).
92. Preeti; Kumar, M.; Sharma, H.K. A GAN-Based Model of Deepfake Detection in Social Media. *Procedia Comput. Sci.* **2023**, *218*, 2153–2162. [CrossRef]
93. CelebA Dataset. Available online: https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html (accessed on 21 January 2024).
94. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* **2015**, arXiv:1511.06434.
95. Yang, C.; Ding, L.; Chen, Y.; Li, H. Defending against GAN-based Deepfake Attacks via Transformation-aware Adversarial Faces. *arXiv* **2020**, arXiv:2006.07421v1.
96. Nadimpalli, A.V.; Rattani, A. ProActive DeepFake Detection using GAN-based Visible Watermarking. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**. [CrossRef]
97. Giudice, O.; Guarnera, L.; Battiato, S. Fighting deepfakes by detecting GAN DCT anomalies. *J. Imaging* **2021**, *7*, 128. [CrossRef]