applied sciences

MDPI

Review

From Tweets to Threats: A Survey of Cybersecurity Threat Detection Challenges, AI-Based Solutions and Potential Opportunities in X

Omar Alsodi [1,*], Xujuan Zhou [2], Raj Gururajan [2,3], Anup Shrestha [2] and Eyad Btoush [2]

[1] Business Intelligence Department, School of Business, Al-Zaytoonah University of Jordan, Amman 11733, Jordan
[2] School of Business, University of Southern Queensland (UniSQ), Springfield, QLD 4300, Australia; xujuan.zhou@unisq.edu.au (X.Z.); raj.gururajan@unisq.edu.au (R.G.); anup.shrestha@unisq.edu.au (A.S.); eyadabdellatif.a.q.marazqahbtoush@unisq.edu.au (E.B.)
[3] School of Computing, SRM Institute of Science and Technology, Chennai 603203, India
* Correspondence: omar.alsodi@zuj.edu.jo

Abstract: The pervasive use of social media platforms, such as X (formerly Twitter), has become a part of our daily lives, simultaneously increasing the threat of cyber attacks. To address this risk, numerous studies have explored methods to detect and predict cyber attacks by analyzing X data. This study specifically examines the application of AI techniques for predicting potential cyber threats on X. DeepNN consistently outperforms competing methods in terms of overall and average figure of merit. While character-level feature extraction methods are abundant, we contend that a semantic focus is more beneficial for this stage of the process. The findings indicate that current studies often lack comprehensive evaluations of critical aspects such as prediction scope, types of cybersecurity threats, feature extraction techniques, algorithm complexity, information summarization levels, scalability over time, and performance measurements. This review primarily focuses on identifying AI methods used to detect cyber threats on X and investigates existing gaps and trends in this area. Notably, over the past few years, limited review articles have been published on detecting cyber threats on X, especially those concentrating on recent journal articles rather than conference papers.

Keywords: social media; cybersecurity; survey; artificial intelligence; security and privacy; natural language processing; cyber threat detection; X

Academic Editor: Stefan Fischer

Received: 28 November 2024
Revised: 31 December 2024
Accepted: 9 January 2025
Published: 2 April 2025

Citation: Alsodi, O.; Zhou, X.; Gururajan, R.; Shrestha, A.; Btoush, E. From Tweets to Threats: A Survey of Cybersecurity Threat Detection Challenges, AI-Based Solutions and Potential Opportunities in X. *Appl. Sci.* **2025**, *15*, 3898. https://doi.org/10.3390/app15073898

1. Introduction

The integration of artificial intelligence (AI) has dramatically reshaped the cybersecurity landscape, introducing both powerful defenses and potent threats. While AI excels at identifying anomalies, authenticating users, and responding to incidents, malicious actors are exploiting its capabilities to create increasingly sophisticated attacks. This complex interplay between AI and human adversaries has generated a rapidly evolving threat environment. AI-powered attacks, capable of bypassing traditional defenses, pose a significant risk to organizations. Effective countermeasures require a multifaceted approach that combines advanced threat intelligence, adaptable defenses, and a strong ethical framework. Leveraging AI defensively can enhance threat detection, automate responses, and augment human analysts. However, challenges such as algorithmic bias, data privacy concern, and the potential for AI-driven attacks necessitate careful risk management. To fully realize AI's potential in cybersecurity, organizations must prioritize regulatory compliance, industry

standards, and collaboration. Investing in cybersecurity education and training is crucial to develop a skilled workforce capable of addressing emerging threats. By bridging the gap between theory and practice, we can effectively mitigate AI-related risks and build a more resilient digital ecosystem [1]. Cybersecurity threats have become a major concern for social media platforms in recent years. This coincides with a booming cybersecurity market, which has grown approximately 35 fold in the past decade. In 2019, global cybersecurity spending reached USD 40.8 billion, rising steadily to USD 71.1 billion by 2022 [2]. As of 2023, spending topped USD 80 billion, and forecasts predict that it will exceed USD 87 billion in 2024. This surge in cybersecurity spending reflects the increasing threat landscape. The digital economy's growth has unfortunately been accompanied by a rise in digital crime. The explosion of online and social media applications has created more opportunities for attackers, leading to data breaches that endanger both users and social media platforms. At the current rate of growth, the financial damage caused by cyber attacks is projected to reach nearly USD 10.5 trillion annually by 2025, marking a 3-fold increase from the levels recorded in 2015 [3]. Global cybersecurity spending from 2017 to 2024 is illustrated in Figure 1 [2].

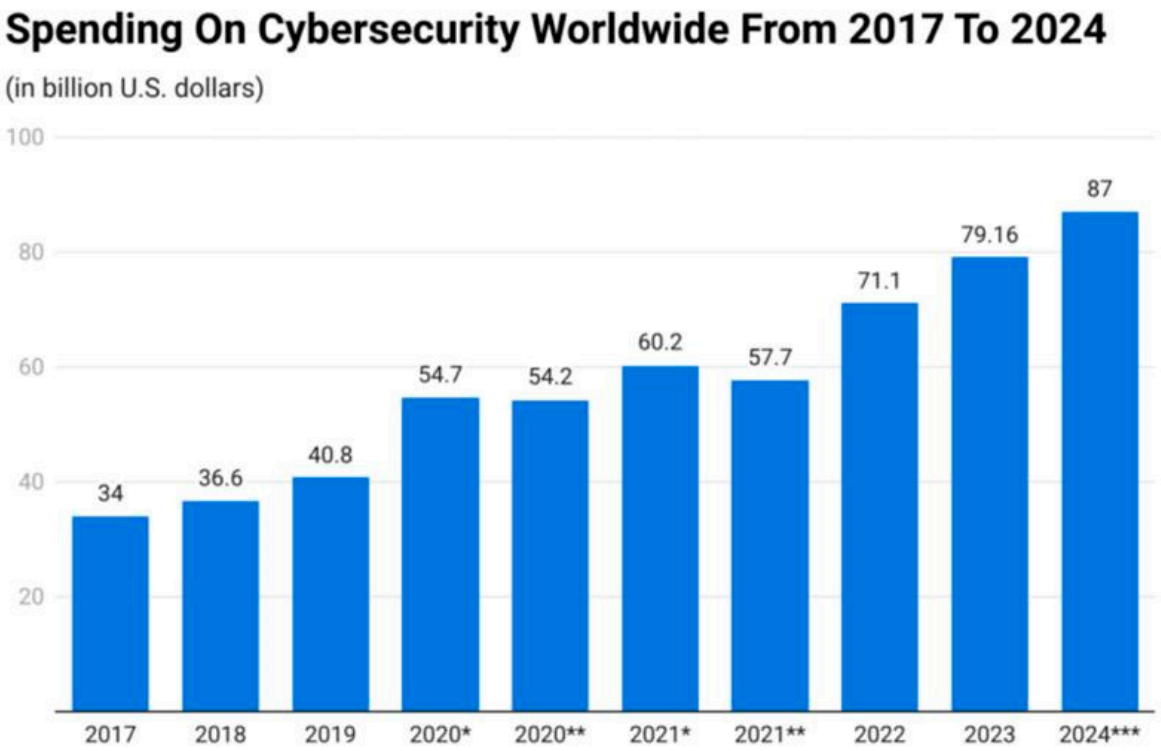

Figure 1. Global cybersecurity spending from 2017 to 2024. * Preliminary data, ** Estimated values, *** Forecasted data.

The surge of online social media platforms like X, Facebook, and TikTok reflects our evolving relationship with data sharing in the digital age. However, this convenience comes with a growing risk: cyber threats. Cyber threats involve criminals using technology to steal sensitive data, like users' information, through cyber attacks. These stolen data can then be used to perform unauthorized activities online. Lost, stolen, or skimmed information can all be vulnerabilities for fraudsters. As the volume of social media platforms continues to climb, so does the threat of cyber threats, posing a serious challenge for both individuals and the social media platforms [4]. X comprises online services that enable users to establish a public or semi-public profile and connect with a list of other users to view and share their profiles and content. The association of X links differs from one service to another [5].

There is a growing range of X with several common features [6]. Social networks are online platforms where users can: (1) Create a public or partially public profile with limitations set by the platform, (2) build a list of connections with other users they know, and (3) browse their connections and connections of others to navigate the social network.

X report different cybersecurity attacks against them that aim to steal the identity of users or undermine the privacy and trust of the network. These threats include activities such as hijacking, identity theft, spamming, social phishing, malware attacks, face image retrieval and analysis, impersonation, fake requests, and Sybil attacks. Attackers, also known as hackers, carry out attacks on X with a wide range of motivations that include political, emotional, financial, entertainment, ideological, personal, cyber warfare, and commercial purposes. As cyber threats increase security risks, numerous researchers and security firms have been developing several solutions. Watermarking [7], Steganalysis and digital oblivion [8] are some of the solutions for protecting X users against threats from compromised multimedia data. Likewise, traditional solutions such as spam detection [9] and phishing detection mitigate the conventional risks. There are also some established security solutions such as mechanisms for authentication [10] and privacy settings [11] as well as commercial solutions such as minor monitoring and social protection applications that offer safeguards against cyber threats in X. Thus, the traditional information security solutions that focus on heuristics and digital signatures are predominantly static and do not offer full protection against the dynamic nature of the new generation of cybersecurity threats that are more evasive and resilient, [12]. However, existing cybersecurity solutions are not robust in detecting cybersecurity threats on X. There are two primary reasons for this problem. Firstly, since the tweets are limited to 140 characters and the writing patterns of people are flexible, the meaning and context of words are also used and are varied [13]. Secondly, there are many diverse and confounding advertisement tweets and people misuse hashtags in their posts to get attention. For these reasons, it is extremely difficult to detect cybersecurity threats from tweets [14]. Cybersecurity threats have become a critical concern in recent years with the growing popularity of social networks. X-based event detection has become a popular method of communicating such threats, and researchers have been using X as an extensive database for event analysis and extraction. Various techniques have been proposed for the detection of cybersecurity threats in X, focusing on attributes, frequency, and multimodal X hashtags. However, the current studies lack comprehensive evaluations of critical factors such as prediction scope, type of cybersecurity threats, feature extraction technique, algorithm complexity, information summarization level, scalability over time, and performance measurements.

This paper focuses mostly on finding AI methods used to detect cyber threats on X. Furthermore, we aim to investigate the gaps and trends in this area. Over the last few years, limited review articles have been published on detecting cyber threats on X. This review looks at the detection of cyber threats on X using machine and deep learning techniques. Further, unlike other analyses that include conference articles, our paper contains recent journal articles.

This study gives important background information on threats from cyber targeting X. First, an overview of cybersecurity threats in X is provided, followed by an explanation of the specific challenges and threats encountered on this platform. The incentives driving cyber threats on X are then examined, followed by a description of the methodology used in this paper. The research then investigates cyber threat solutions and analyzes the most recent ones. Following that, a gap analysis of existing research and recommendations for future approaches are presented. The limitations of the survey are also discussed. This paper closes with the conclusion.

2. Cybersecurity Challenges and Threats in X

Cybersecurity is a tool to detect unwanted access to the property of individuals and organizations [15]. The cybersecurity community has established the field of Cyber Threat Intelligence (CTI). Cyber Threat Intelligence (CTI) has been receiving increasing attention from both academic and CTI researchers in security operating centers and security service providers as a component of cybersecurity [16]. The primary objective of CTI is to develop a knowledge advantage over cyber threat actors. At the tactical and operational levels, CTI expedites early detection of malicious behaviors, preferably before a malicious actor gains a foothold in the network. On a strategic level, CTI provides sense-making and insight into the relevant threat environment to decision makers. Effectively, CTI is the civilian, private-sector alternative to defensive counter-intelligence executed by the established Intelligence Community (IC) [17].

X, with its various features such as tweets, video and image sharing, and e-commerce capabilities, has become an integral aspect of the daily routines of a vast number of internet users. However, this widespread utilization of the platform also exposes individuals to a plethora of cyber threats and security concerns. The following section will outline these potential threats. As illustrated in Figure 2, there are several categories of security threats on X.

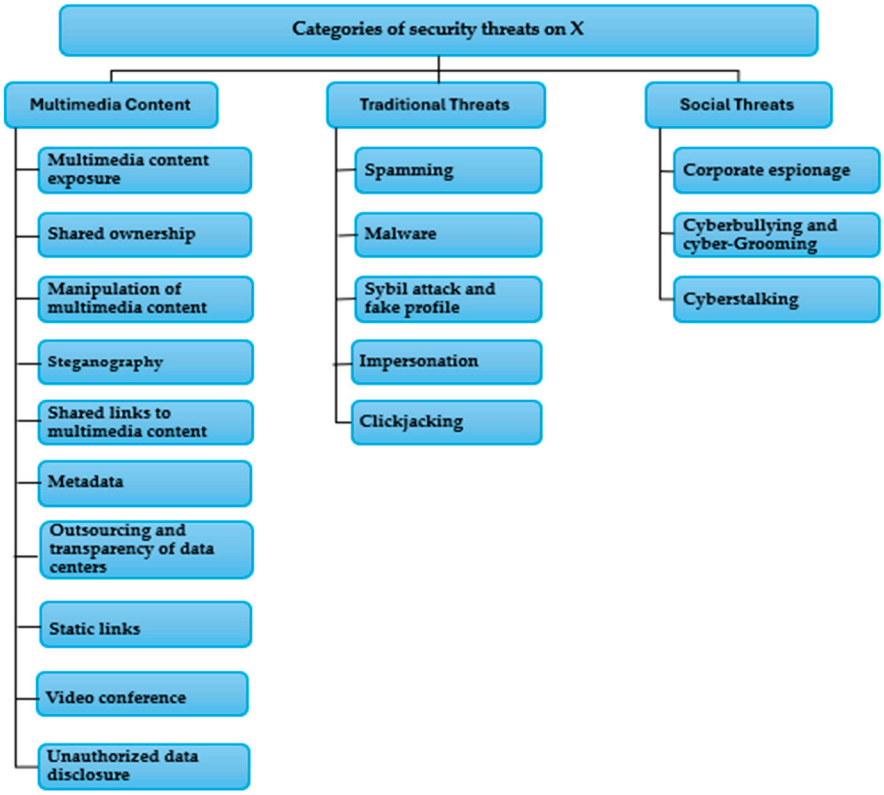

Figure 2. Categories of security threats on X, adapted from [18].

As a leading social media platform with a massive user base and rapid information exchange, X is a prime target for cyber criminals. This section delves into the various cyber threats that plague the platform. X has become a breeding ground for a multitude of cyber attacks, including.

2.1. Security Threats in X

Cybersecurity threats occur more frequently with the popularity of today's use of X. Consequently, these threats may seriously impact the lives of individuals and cause social

and financial unrest. Researchers have been using X at least since 2010 as an extensive, publicly available database for analyzing and extracting cybersecurity threats, the security threats on X are as follows.

2.1.1. Multimedia Content Threats

X allows users to share various forms of data, including multimedia content, which has been improved by the integration of high-definition videos and images. However, multimedia search technologies, such as geotagging and facial recognition, can increase the potential for illegal use of shared data, putting sensitive user information at risk. This section focuses on the multimedia threats that attackers could exploit to obtain sensitive user information from multimedia data shared on X [18].

- Multimedia content exposure

Individuals using social media platforms, like X, are generally cautious about sharing text-based information, such as their identity and home address. However, they often overlook the risks associated with sharing multimedia data. For example, posting a picture of their home can help intruders locate their address. Updates about their whereabouts can indicate an unoccupied home, increasing the risk of burglary. Photos can also reveal the user's current location, aiding potential intruders. Additionally, shared images may draw attention to valuable assets, and posting photos or videos without consent can compromise others' privacy. Technological advances, like facial and voice recognition, further exacerbate these privacy concerns by identifying individuals without their knowledge [19].

- Shared ownership

Multimedia data shared on X may relate to multiple users [20,21]. An illustration of this scenario would be two individuals who are friends attending an event together and subsequently capturing an image together. Subsequently, one of the friends may choose to upload the image to the X platform, without obtaining the consent of the other friend. This may result in the exposure of the other friend's privacy, as the image belongs to both individuals. It is important to note that the preferred privacy settings for multimedia data that pertain to multiple users are determined by a single individual, as opposed to being determined by the intersection of the privacy settings of each individual user, which would be a logical approach [4].

- Manipulation of multimedia content

X offers a medium for users to disseminate and access a plethora of multimedia content. However, the veracity and integrity of this content may be compromised by malicious actors who employ various digital tools to manipulate and distort multimedia data. This can lead to the unauthorized alteration of personal images, resulting in potential harm or defamation of legitimate users [18].

- Steganography

Steganography involves concealing data within other media forms and has gained popularity due to technological advancements. It was discovered that X users employ this technique to hide messages within images, demonstrating its feasibility and low technical complexity. However, this capability can be misused for malicious purposes, such as spreading disinformation and harming the platform's reputation. Innocent users might unknowingly interact with harmful content, risking association with criminal activities. For instance, malicious actors might embed harmful code in an image on X, which is then unknowingly downloaded by users [22].

- Shared links to multimedia content

The variety of multimedia formats, like JPEG and PNG, complicates creating a universal framework to support them all. Some formats are prone to attacks or require manual verification, such as interactive flash videos. Social media platforms, like X, generally support a limited range of formats; for example, X allows sharing pictures only as JPEG or PNG and does not widely support animated GIFs. Users can share unsupported multimedia by posting links, but this feature can be exploited by malicious actors. They might replace the link's content with harmful material, risking malware installation or confidential information theft for the user [23].

- Metadata

Metadata on multimedia content, such as identities and locations, are valuable but can expose users to risks. Geolocation tags in images, often added by modern mobile phones, can reveal personal details like religious beliefs or health conditions, posing potential dangers [24]. Different platforms handle multimedia metadata differently. Facebook removes all metadata before uploading images, while Google+ retains all except GPS coordinates. Flickr, by default, shares GPS coordinates to display images from the same location [25].

- Outsourcing and transparency of data centers

The transparency of stored media on social networks (SNs) poses significant privacy risks for users in two main ways. First, multimedia data on these platforms are often unencrypted, allowing malicious users to access it directly if they obtain a link. Second, data stored on SNs can be viewed by service providers. While major networks like LinkedIn and Facebook operate their own data centers, smaller networks typically rely on third-party cloud storage, which can lead to increased privacy and security concerns despite cost and scalability benefits [26]. End users may trust X, but they struggle to trust third parties with their data. User information can be shared with government agencies for law enforcement and utilized by merchants for marketing [14].

- Static links

Generally, most X end users use static links to share mixed-media data. This is because these links provide an efficient and optimal method of data distribution. However, sharing static links compromises user privacy and can open many opportunities for attacks. When a user shares an image static link with a group of users of their choice, any member of the group can access the image and share it without the permission of the image owner. Members can also copy and paste the link to share the image outside of social media [18].

- Tagging-link ability from shared multimedia data

X has a feature where you can tag multimedia content, such as videos and images, to increase interaction between users and make searching easier. People can label their own content and add more details, but this can also be a threat to their privacy. For instance, some X users may not want to share their own photos, but a friend could tag their photo to reveal their identity [24]. The primary concern is that tagging can connect an individual who does not have a X account and does not wish to reveal any personal details on the platform [21]. Additionally, a spammer or an individual with malicious intent can tag a substantial number of individuals in a single post, such as an image or video, to disseminate harmful content to a wide audience with minimal effort [27].

- Unauthorized data disclosure

X offers its users the ability to share data. Sharing data involves making it available to a specific group of users. However, there is a risk that one of the members of the group may disclose the shared information [28]. This type of disclosure is often considered illegal

as it can be manipulated. The same goes for multimedia data, such as pictures. When a user shares a picture with a group, any member of the group can download it and change the privacy settings, potentially causing the picture to be publicly shared even though the original uploader only wanted it to be seen by a certain group of people.

- Video conference

Today, X offers both text messaging and video conferencing capabilities. The added benefit of video conferencing is that it allows for greater interaction between users. However, this also opens the possibility of more sensitive information being shared. A malicious user can access the video stream by exploiting any vulnerabilities in the communication infrastructure [29]. Additionally, someone participating in the video conference can record it and use it to blackmail or manipulate others. The attacker may also be able to access the webcam of the target by utilizing malware and taking advantage of weaknesses in the communication protocols.

2.1.2. Traditional Threats

In the context of X, there are specific types of traditional threats that involve utilizing various attack methods, such as phishing and malware, to acquire a user's personal details. This information can provide a significant advantage for the attacker, as they can obtain sensitive information such as social security numbers, passwords, and bank information. With this information, the attacker can carry out further crimes such as phishing and identity theft [30]. This section outlines the different traditional threats that can be employed by attackers to access a user's personal information.

- Spamming

Spam attack attackers flood internet users with unsolicited messages (spam). On X, this kind of attack appears to be more successful than traditional spam attacks that use email to spread spam. This is because the social connections between X users can be easily abused. Target users can easily be convinced to read spam information and trust it to be safe. Here, the attacker can somehow obtain communication details about the user and send spam or junk data. Obtaining communication details is not too difficult and can be extracted from legitimate user profiles. A large amount of spam emails sent causes network congestion and the cost of sending emails is mainly borne by the provider of the service and in some cases by the user [31].

- Malware

This is harmful software made up of Trojan horses, viruses, and worms. X operates by connecting different users' systems. As a result, malware can easily spread from one user's system to another through these connections [32]. X lacks the necessary tools to identify if a URL is dangerous or not. Dangerous URLs can steer users to fake websites which can then transmit malware to their computers and steal their confidential information. Researchers looked at the spread of malware on X and determined which factors played a role in its spread [33]. These factors include features of the social network graph such as the number of nodes, number of connections, highest degree, average shortest distance, and longest distance. The researchers also explained how each factor affects the rate at which malware spreads on X [34].

- Sybil attack and fake profiles

In a Sybil attack, attackers generate a significant number of fake identities to gain an advantage in distributed and peer-to-peer systems. This type of attack poses a significant threat to X security as it has many users connected as peers in a peer-to-peer network, allowing one entity to control multiple fake identities. By utilizing these fake identities, attackers

can override legitimate users and manipulate reputation values, corrupt information, and outvote legal X users, such as by voting an account as the "best" [35].

- Impersonation

The goal of the attacker is to construct a false profile with the intention of pretending to be a real individual. This type of attack heavily relies on the authentication procedures that users encounter when creating a new account. Such attacks can have severe consequences for the person being impersonated [31].

- Clickjacking

This is a growing threat to X where attackers conceal harmful software behind the sensitive user interfaces or buttons to steal clicks from customers for malicious purposes. Clickjacking has various forms, but the most well known are Likejacking and Cursorjacking. In Likejacking, the attacker embeds malicious code scripts with X's "Retweet button" that appears on the user's profile. Cursorjacking employs the user interface redressing technique to change the position of the cursor, where the attacker replaces the actual cursor with a fake one to direct the user to a malicious website [36].

- Social phishing

This type of attack involves the attacker attempting to obtain confidential information from a target by using a fake website that appears authentic or by pretending to be someone the target knows. The severity of these attacks can be significantly reduced if the target is informed and cautious when reviewing information received [30].

- Hijacking

Gaining control over another person's profile is referred to as hijacking. The attacker succeeds in this if they can guess or obtain the login password for the account. Choosing weak passwords is not recommended as it increases the risk of hijacking. These passwords can easily be acquired through dictionary attacks. To prevent this, it is best to use strong passwords and change them frequently [31].

2.1.3. Social Threats

Regarding online social threats, attackers can utilize the social relationship aspect of X to maliciously engage with different user groups, such as minors and corporate employees. For example, an attacker may manipulate minors by expressing compassion and offering online gifts or money. Their reasons for doing so can range from blackmail, distributing pornography, conducting cyberbullying, and espionage [4]. In this section, we outline the various social dangers that take advantage of different online social relationships for different motives.

- Corporate espionage

Corporate espionage can employ automated social engineering tactics through X. By utilizing X as a tool, a social engineer can obtain valuable information, such as the job title, email, and complete name of employees, without relying on traditional social engineering methods and infiltrating the company. A study by [36] describes a method of using social networking sites (X) to execute a social engineering attack. They demonstrated that by utilizing X, an attacker can gather information about an employee within a targeted organization in an automated fashion, which can then be utilized for a successful social engineering attack [37].

- Cyberbullying and cyber-grooming

Cyberbullying is the repeated online harassment of an individual, while cyber-grooming involves an adult forming an emotional bond with a child to facilitate sexual

abuse. Children are especially susceptible to these online threats due to their age [38]. Teenagers facing cyberbullying are at an increased risk of depression. Online predators often exploit this vulnerability by posing as caring individuals, luring victims with gifts and financial incentives. Security experts report that such predators have targeted thousands of students globally through deceitful practices. A notable case is that of Megan Meier, whose tragic suicide highlighted the severe consequences of cyberbullying, as the perpetrator created a fake online profile for manipulation and cyber-grooming [39].

- Cyberstalking

X users have the option to reveal their personal details such as contact information, home address, location, and schedule on their X profile. However, this information can be vulnerable to exploitation by malicious individuals for cyberstalking purposes. For example, an attacker can blackmail their victim through phone calls or instant messages on X. Additionally, users often share location information through their photos, which attackers can gather and use for harmful cyberstalking attacks. Researchers reviewed the effects of cyberstalking on German X users on StudiVZ. They emphasized that cyberstalking could harm the mental well-being of X users and should be regarded as a significant danger to ensure a safe and secure environment on the platform.

3. Motivations of the Cyber Threats on X

Hackers have increasingly become a major concern for X users, executing various attacks driven by different motivations. These can include revenge, financial gain, entertainment, or participation in hacktivist movements that protest specific issues. Some hackers also engage in espionage or cyber warfare for political or military reasons. Regardless of their intentions, these attacks can have serious repercussions for individuals and organizations. Therefore, it is crucial for X users to understand these motivations and take proactive measures to safeguard their online presence [4].

❖ Financial benefits

Financial benefits are the primary motivation behind cyber attacks on X. These attacks are carried out by cyber criminals who aim to acquire sensitive information related to the bank accounts of users [40]. The malicious access of these accounts allows the perpetrators to steal money and financial assets from the victims [41]. Additionally, business-related information can also be targeted in these attacks, with the intention of profiting from the information by rival companies. The ease of access to large amounts of personal and financial information on X makes it a prime target for cyber criminals looking to make quick and easy financial gains [42].

❖ Entertainment

Entertainment can come in many forms and for some hackers, it lies in the excitement of hacking on social media. These individuals are driven by the thrill of showcasing their hacking skills to their peers and gaining recognition in the hacking community. They do not have any financial or political motives behind their actions, but simply do it for the enjoyment of the challenge. As the saying goes, some people just find pleasure in causing chaos and disruption. For these hackers, hacking is a form of entertainment that allows them to express their technical abilities and gain a sense of notoriety among their peers [4].

❖ Cyber spying

Cyber espionage refers to the act of obtaining private information without the permission of the owner using hacking techniques and malicious software. This type of espionage is becoming increasingly prevalent on social media, where individuals, competitors, and even foreign governments are targeting confidential information. This can range from

personal data to sensitive business information and can have serious consequences for those affected. The rise of cyber espionage highlights the importance of taking necessary precautions to protect personal and business information online [43].

❖ Expertise for the job

The demand for expertise in the fields of cybersecurity and hacking is at an all-time high, as many IT experts lack these specific skill sets. The job market for these positions is extremely competitive, as organizations are eager to hire individuals who can help them evaluate their security and protect against cyber criminals. Having a specialist on their team allows companies to think and operate in the same way as the criminals, giving them a better chance at beating them. The need for these experts is crucial in today's world, as cyber threats continue to grow and evolve [44].

❖ Cyber warfare

Cyber warfare is a new form of conflict that is fought using technology and the internet. Cyber warfare is a politically motivated attack on information and information systems, mainly targeting government websites. The goal of these attacks is to disrupt the communication and financial stability of the targeted country and to cause improper functioning of its government. Unlike traditional warfare, cyber warfare is fought from the comfort of a room rather than on the front lines. The use of social media has made it easier for individuals or groups to launch these attacks, making it a serious threat to national security [45].

❖ Revenge/Feelings

Revenge and emotions can drive individuals to engage in cyber attacks on X. Whether it is a dissatisfied customer or an unhappy employee, the desire for revenge can lead to the destruction of an organization's reputation. These hackers aim to cause chaos and frustration by blocking services and leaving legitimate users without access. The impact of such attacks can be devastating, causing significant financial loss to the victim organization. It is important to recognize the power of emotions and the potential consequences they can have in the digital world [46].

❖ Hacktivism

Hacktivism is a form of activism that utilizes technology to achieve political and social goals. The main objectives of hacktivism include promoting free speech, protecting human rights, and advancing information ethics. This type of activism involves publishing the views and aims of a political community or religious group and staging protests to support their beliefs. However, it can also involve vandalism of websites with political or religious messages. Hacktivism is a unique form of activism that combines technology and activism to bring attention to important political and social issues [47]. Figure 3 illustrates the impact and motivation of cyber threats on X.

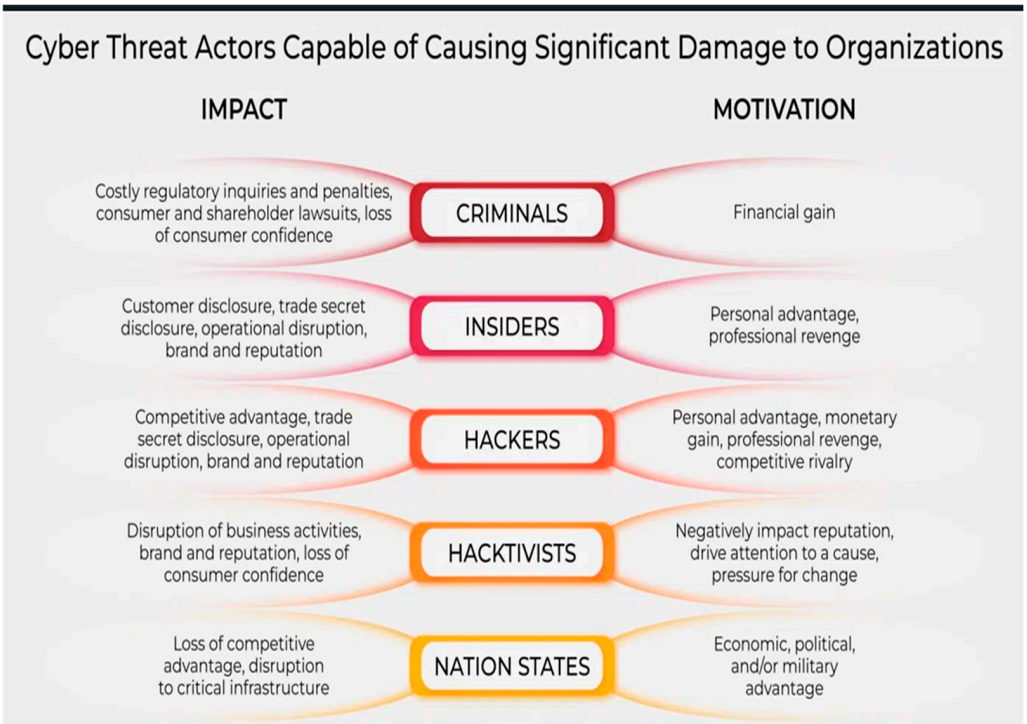

Figure 3. Example of the cyber threats impact and motivation on X.

4. Survey Methodology

- Research Questions

  (RQ 1): What are the cyber threats present on X, and what motivates these threats?
  (RQ 2): What AI-based solutions can be employed to address cyber threats on X?
  (RQ 3): What potential opportunities can be explored through these solutions?

- Objectives

One area of concern is the use of AI in cyber attacks on X. Cyber criminals are increasingly using AI-based solutions to carry out highly sophisticated and difficult-to-detect attacks. To answer our research questions, this paper examines the effectiveness of current cybersecurity measures in detecting and preventing such attacks and assesses the limitations of traditional cybersecurity methods. Additionally, this paper discusses the potential benefits of using AI-based solutions to combat X cyber threats. Traditional cybersecurity techniques are often inadequate when it comes to addressing the ever-evolving nature of cyber threats. AI-based solutions, on the other hand, can quickly learn and adapt to new threats, allowing them to respond more effectively and efficiently to potential attacks. This paper also explores the potential for AI-based solutions to address X's cybersecurity challenges. By highlighting the unique features and vulnerabilities of this X platform, this paper aims to contribute to our understanding of cybersecurity issues and facilitate the development of effective solutions to address them. We aim to promote the use of AI-based solutions to combat social media-based threats and improve the overall security on X.

5. AI-Based Cyber Threat Solutions in X

AI algorithms play a crucial role in the pattern recognition capabilities of machine learning, which can be divided into two main categories: supervised and unsupervised algorithms. Supervised algorithms use labeled data to predict image classes, including parametric models like Support Vector Machines and non-parametric methods such as k-Nearest Neighbors. In contrast, unsupervised algorithms analyze unlabeled data to

identify patterns through clustering and dimensionality reduction techniques. Choosing the right algorithm depends on factors like accuracy, scalability, and the specific problem being addressed. Despite initial skepticism, the advantages of AI have gained acceptance, particularly in the realms of machine learning and deep learning, with computer vision being a significant area of focus. Understanding these relationships is key to appreciating advancements in machine vision [48].

X has become a significant platform for the dissemination of information, including cyber threats. This presents both challenges and opportunities for cybersecurity researchers. Machine and deep learning offer powerful tools to analyze this vast and dynamic data stream, enabling more effective threat detection, response, and prevention [49]. Machine learning, deep learning and Ensemble Learning for cybersecurity threat detection are explored in the following sections.

5.1. Machine Learning (ML)

Machine learning (ML) is a branch of artificial intelligence (AI) and computer science that focuses on the using data and algorithms to enable AI to imitate the way that humans learn, gradually improving its accuracy [50]. Machine learning can create an effective model automatically based on initial training data. The motivation for this approach is the availability of the appropriate training data, or it can be obtained at least more conveniently compared to the effort required to define the model manually [51]. The versatility of ML has led to its widespread application across diverse sectors, including healthcare, finance, natural language processing, and autonomous vehicles, as it can process large datasets and reveal insights that may be difficult for humans to discern. The scalability and adaptability of ML models make them highly valuable for addressing complex real-world problems, fueling innovation across industries [52]. Additionally, advancements in computational power and access to big data have accelerated the development of more advanced ML techniques, such as deep learning, which emulates neural networks in the human brain to analyze vast amounts of data with greater accuracy [53]. As ML continues to evolve, its potential to transform industries and solve pressing challenges expands rapidly.

This section delves into the three primary methodologies in this field: supervised, semi-supervised, and unsupervised learning. This section provides a contextual background and a comprehensive analysis of key research within each category.

1. Supervised Learning

Supervised learning is a machine learning system that can learn from the training data. Training data consist of pairs of input objects (usually vectors) and outputs. The output of the function can be assumed to be a continuous value (called regression) or a class mark of the input object (called classification). The task of the supervised learner is to predict the value of the function of any valid input object after observing several training examples (i.e., pairs of input and target output). In order to do so, the learner must "reasonably" generalize from the data given to unseen circumstances [54]. In other words, in terms of predictor characteristics, the goal of supervised learning is to create a concise model for the distribution of class labels [55]. Supervised learning is used, in particular, as a predictive mechanism in which a portion of the data is learned (or otherwise known as a training set), while another portion is used to test a trained model (Cross-validation) and the remainder will be used to determine the accuracy and effectiveness of the forecast [56]. Cyber threat analysis primarily relies on two types of features: behavioral and content. Behavioral features focus on user metadata, actions, and interactions, without deep content analysis. They examine factors like timestamps and basic text counts. The content features delve into the textual content itself to differentiate bots from real users.

- **Behavior based**

The well-known BotOrNot [57] is an off-the-shelf system that leverages more than one thousand features to discriminate bots. BotOrNot measures the 'botness' of an X account. The authors expanded on their previous work [56] by retraining the model on a new dataset [58] and disclosing their feature engineering process [59]. They developed a feature set inspired by the DARPA competition [60] to distinguish between normal and bot accounts. Ref. [61] identified and evaluated the importance of features for Sybil detection on X, finding Random Forest to be the most effective classifier. Ref. [62] combined Support Vector Machines and neural networks (SVM-NN) to detect fake accounts and bots, reducing the feature set from to improve efficiency [63]. Ref. [64] developed a method to calculate a 'botScore' for X accounts, like the BotOrNot botness score. They identified ten user profile attributes and tweet patterns to feed into their BotClassifier, a supervised classification algorithm. Compared to Naive Bayes, their model demonstrated superior performance in distinguishing between human and bot accounts. CATS by [65] uses a clever approach to spot X spam bots. By analyzing just 5 tweets per user, they combine entropy, spammer behavior, and a blacklist of spammy URLs. This helps them accurately identify spam accounts. The CATS team also introduced 15 new features for better spam detection. They tested different machine learning methods and even grouped spammers to understand how they operate. Ref. [66] conducted an empirical study on the evasion tactics used by social bots. They identified key characteristics of social bots and common evasion techniques, and subsequently proposed a detection method that incorporated nine novel features alongside existing ones. Their approach was evaluated across multiple social media platforms. Ref. [67] introduced a hybrid approach combining human judgment and machine learning to identify X bots. This semi-automated method prioritizes precision, making it suitable for creating large, high-quality datasets for bot detection models. Ref. [68] developed a novel approach to classifying X accounts by stratifying users based on account popularity. This strategy, centered on user social status, allowed them to identify distinct feature sets effective for distinguishing between human and automated accounts within each popularity tier. While their study focused on general account classification rather than specifically detecting malicious bots, the methodology, features, and dataset generated could serve as valuable foundations for future bot detection research. Similarly, ref. [69] introduced a refined set of features focusing on user interaction levels and engagement. They combined these features with existing ones to detect X bots using deep learning. In a similar vein, ref. [70] identified bots in marketing campaigns by analyzing user interactions on X. They compared various classifiers and found back-propagation neural networks to be most effective with their feature set. Pattern recognition techniques have been applied to classify X accounts. Ref. [71] developed a model to categorize accounts as human, bot (spam bot), or cyborg. Their approach involved analyzing account behavior through entropy calculations for tweet timing patterns, machine learning for text-based spam detection, and statistical analysis of account properties. A decision-making component combined these analyses for final classification. In subsequent work [72], the model was refined with enhanced components and evaluated on a larger dataset. Ref. [73] developed a method to identify automated fake X profiles by analyzing multiple profile attributes, screen name patterns, and tweet posting times. While their model exhibited high precision in detecting fake accounts, its recall was relatively low. Nonetheless, due to its exceptional accuracy, the authors propose using it as a baseline or starting point for more complex graph-based detection methods. Ref. [74] developed an algorithmic approach to identify distinct behavioral patterns between real and fake X users. Their method focuses on extracting Relaxed Functional Dependencies to differentiate between human and bot accounts. The researchers posit that the complex patterns exhibited by humans are inherently difficult for

bots to replicate. Ref. [75] introduced a proactive method to identify Sybil accounts during their creation. By comparing private user data and images, their framework can prevent these fraudulent accounts from being established. Additionally, ref. [76] developed a model that represents social media users based on their behavior and posting patterns. This model, utilizing a CNN-LSTM algorithm, was employed to distinguish between human and bot accounts on X.

- **Content based**

Numerous studies have focused on content analysis and textual information to detect X bots. For instance, ref. [77] employed deep learning to identify bots using a single tweet and six account features. They addressed dataset limitations by applying oversampling techniques to a small training set. Similarly, ref. [78] leveraged tweet similarity to detect social bots, assuming similar botmaster objectives and technological constraints lead to comparable tweet content. Ref. [79] employed a CNN-LSTM algorithm on tweet content and metadata to identify evasive spam bots. Given the role of bots in misinformation spread, ref. [80] proposed using topic analysis for bot detection. Their Boost OR algorithm optimized F1-score by balancing precision and recall. Notably, they introduced two publicly available labeled X datasets. Ref. [81] classified X accounts into human, bot, and cyborg using pattern recognition and a wavelet-based approach. Random Forest outperformed Multilayer Perceptron, especially in the binary human/non-human classification. Ref. [82] extended this work to distinguish between humans, legitimate bots, and malicious bots. Assuming similar patterns in spam generated by the same botmaster, ref. [83] developed an iterative model to detect X spam and spam bots based on tweet similarity and closeness to known spam. While not directly focused on bot detection, ref. [84] classified X users into person and non-person. Their first step involved using Xati [85] to identify bots based on tweet properties like inter-tweet delay, spam detection, near-duplicate tweets, Klout score, and tweeting device. Ref. [86] hypothesized that sentiment differences could distinguish humans from bots. Ref. [86] introduced sentiment-based features alongside other tweet and user characteristics. Ref. [87] also used sentiment analysis with a Contrast Pattern-Based classifier. Ref. [88] employed content and metadata information to detect social spam bots.

In a different approach, ref. [89] adopted a different approach by focusing on usernames rather than user posts. They categorized usernames as either random or non-random, creating a dataset of 235,000 X accounts with random usernames, which they labeled as automated. An analysis of a 100-account sample from this dataset led them to conclude that it is accurate and diverse, making it a valuable resource for improving bot detection on social media.

2. Unsupervised Learning

Some of these techniques does not require training data. They are based, as alternatives, on two fundamental assumptions. Firstly, they assume that daily traffic is the majority of network connections and that only a very small percentage of traffic is abnormal. Secondly, malicious traffic is calculated to be statistically different from regular traffic [50]. According to these two assumptions, daily traffic is typically presumed to be data groups of similar instances, although occasionally instances that differ significantly from most instances are considered to be malicious [90]. Datasets provided as machine learning input in unsupervised learning are not labelled in any way that defines the correct or incorrect outcome. Instead, the result may achieve a larger desired target, be measured on the ability to find something readily discernible by humans, or provide a nuanced application of the statistical function to obtain the intended value [55]. Similar to supervised learning articles, those employing unsupervised methods are categorized into behavior-based and content-based approaches. A review of these unsupervised techniques follows:

- **Behavior based**

Several models have been proposed to detect social media bots using unsupervised machine learning techniques. DeBot by [91] identifies bots based on correlated activity patterns. Assuming human users exhibit less correlated behavior over time, DeBot flags accounts tweeting frequently (at least 40 tweets/hour) with high activity correlation as potential bots. Ref. [92] introduced the Digital DNA model, which analyzes the sequence of online actions to identify bot campaigns. Accounts with similar action sequences (Longest Common String) are classified as potential spam bots. Their subsequent work [93] applied this model in both supervised and unsupervised settings, favoring the latter. They employed a similar approach in their BotWalk model. By constructing vector representations of user features, BotWalk utilizes seed bots and these vectors to identify social bots on X. Seed bots are discovered using DeBot [94], and the model then expands its search to connected users to detect anomalous accounts. The dataset used for this research is publicly accessible.

- **Content based**

To disseminate information effectively, bots typically exhibit openness and content duplication. Ref. [95] exploited these characteristics to identify patterns of similarity and subsequently detect automated X accounts, commonly referred to as Influence bots. By analyzing tweet data, the author discovered emerging patterns among groups of accounts, positing that these patterns alone suffice to classify accounts as automated without requiring additional ground truth verification. Building on this concept, ref. [96] developed a method to detect spam bot campaigns on X by examining patterns in URL shortening services and comparing content similarity between tweets. In a subsequent study, ref. [97] designed a system capable of identifying spam bot campaigns on the X platform. The system identifies groups of accounts sharing identical tweets by monitoring top trending URLs on X's real-time stream. Accounts within these groups are flagged as potential bots if they exhibit similar recent tweeting behavior. A classifier is then employed to distinguish spam bot campaigns based on shared tweet content. Finally, the system links each identified campaign to the email address associated with the URL it promotes. Ref. [98] employed a content-based approach to identify Small- and Medium-Sized Businesses (SMBs) within the BotCamp dataset. Their model capitalized on trending topics to detect social threats campaigns focused on political discourse. By gathering trending hashtags, he model employs DeBot by [91] to identify synchronized bots that exploit popular hashtags. Subsequently, graph-based techniques are used to represent topological relationships between these bots and group them into clusters. A supervised model is then applied to categorize user interactions as either agreeing or disagreeing with specific sentiments. Ultimately, the identified clusters serve as indicators of bot-driven campaigns within political discourse.

3. Reinforcement Learning or Semi-Supervised learning.

Reinforcement Learning is a learning technique dealing with the study of how machines and natural systems, such as humans, learn in the presence of both labelled and unlabeled data. Traditionally, learning has been studied either in the unsupervised paradigm where all data are unlabeled (e.g., clustering, outlier detection) or in the supervised paradigm where all data are labelled (e.g., classification, regression) [99]. In recent years, interest in SSL has increased, especially because of application domains in which unlabeled data such as images, text, and bioinformatics are abundant [100]. The aim of the reinforcement learning approach is to maximize the reward of each change in state by learning the best behavior to be performed in each state [55]. Ref. [101] introduced clickstream sequences as a robust feature to differentiate human users from social bots. By employing semi-supervised clustering, they leveraged the dynamic nature of clickstream

data to unveil subtle behavioral patterns that are challenging for bots to replicate. This approach assumes that clickstream sequences encapsulate both the evolving aspects of user behavior and underlying, consistent characteristics. Leveraging the principle of homophily in social networks, ref. [102] developed SocialBotHunter, a model that detects spam bots on X by analyzing user behavior and interactions. Requiring only a seed set of labeled legitimate users, the model effectively identifies spam accounts.

5.2. Deep Learning (DL)

DL is a type of ML technique that allows machines to learn from their mistakes and comprehend the world as a hierarchy of concepts [103]. DL enables computational models consisting of several layers of processing to learn data representation at multiple abstraction levels. These methods have greatly improved state-of-the-art speech recognition, visual object recognition, object detection, and many other domains such as drug discovery and genomics [52]. The use of DL technology for cybersecurity research and intrusion detection is highly important since most attacks use invasive software families that can be detected and classified [55]. DL is commonly used in pattern recognition. Furthermore the issue of classification, such as text classification and image classification, has also shown efficiency when DL is used [13].

1. Convolutional neural networks (CNNS)

ConvNets is designed to process data that come in the form of multiple arrays, such as a color image consisting of three 2D pixel-intensity arrays in three color channels. There are many data modalities in the form of multiple arrays and 1D for signals. Sequences, like language; 2D images or audio spectrograms; and 3D images, either video or volumetric. The four key ideas behind ConvNets that take advantage of the characteristics of natural signals are: local connections, shared weights, pooling, and the use of multiple layers [52]. Convolutional networks integrate three architectural ideas to ensure a certain degree of transition, size, and distortion invariance: (1) local receptive fields, (2) shared weights (or duplication of weights), and (3) spatial or temporal subsampling [104]. Ref. [105] also suggested CNN to strive for image recognition. The basic idea of CNN is to capture a data function by transferring the kernel, a convolution matrix, to a region in the image. Generally, while neural networks cannot retain spatial information in the image, CNN can maintain it by adding the kernel to each area of the image. In the case of natural language processing (NLP), we can also add the convolutional layer of CNN to the vector space translated from the text corpus. Since each kernel can learn how to insert in a region, i.e., one sentence in the NLP, and capture the semantic and structural features of the sentence, CNN performs well in the text classification. Ref. [14] proposed a multitask learning approach based on the natural language processing technology and ML algorithm of the iterated dilated convolutional neural network (IDCNN) and Bidirectional Long Short-Term Memory (BiLSTM) to establish a highly accurate network model. Their results show that the proposed model operates well to predict cyber hazard incidents from tweets and greatly outperforms a variety of baselines.

2. Graph Convolutional Networks (GCNs)

A Graph Convolutional Network (GCN) is a specialized neural network for processing and analyzing graph-structured data. In graphs, nodes represent entities, and edges represent the relationships between those entities [106]. GCNs have garnered significant attention and popularity due to their effectiveness across various domains where data can be naturally represented as graphs. Their strengths include the ability to excellently represent nodes within a graph through their iterative structure, handle irregular and complex data structures, perform node classification and prediction, and adapt to new

graph contexts and scale efficiently [107]. Ref. [108] proposed a deep learning-based approach for identifying trolls and toxic content on social media. The developed machine learning model detects toxic images by analyzing their embedded text content. The model employs GloVe word embeddings to improve predictive accuracy and incorporates Graph Convolutional Networks (GCNs) to analyze the complex relationships in social media data. While the model demonstrates potential, it faces challenges in precision and recall. The model correctly identifies toxic content in more than 50% of cases but struggles with precision, detecting positive instances less than 50% of the time. Additionally, the recall rate is limited, capturing only 40% of positive cases. The F1-score, which balances precision and recall, is approximately 0.4, suggesting that further improvements are needed for enhanced effectiveness. Ref. [109] presented a graph-based approach for malware detection by constructing a program graph that captures the relationships within a program and developing two enhanced Graph Convolutional Network (GCN) architectures. The first model incorporates label propagation into the GCN to utilize label information, enabling neighborhood aggregation and propagating labels from labeled to unlabeled nodes. The second model introduces residual connections between the original node features and the node representations generated by the GCN layer, improving information flow and mitigating the over-smoothing problem. Experimental results demonstrate that the proposed models significantly outperform baseline GCN and traditional machine learning methods in malware detection, highlighting their effectiveness in program representation learning and malware detection using program graphs. Ref. [110] proposed a deep learning-based framework that analyzes social media across three key domains: users' profiles, the content they share, and the examination of users' unstructured ego-networks. This framework is built on an inductive learning-based graph neural network, enabling a 3D analysis of social media platforms. The proposed model can serve as a benchmark, providing a baseline for future research. Its performance is compared with existing approaches like SVM and LSTM, and experimental results demonstrate its superior performance using the real-world PHEME dataset. Furthermore, the framework can be leveraged as an OSINT (Open-Source Intelligence) tool, contingent on the availability of customized data.

3. Recurrent neural networks (RNNS) and Long Short-Term Memory Networks (LSTM)

A recurrent neural network (RNN) is a recurrent structure where a directed graph along a chain is generated by node associations. This helps the RNN to view time dynamic behavior for a time series applied to natural language processing (NLP). RNNs can use their internal state to process input agreements and may do so only for a limited period of time, i.e., they cannot remember long-term information [110]. In other words, RNN is a neural network that simulates a complex system of discrete time that has an input xt, an output yt, and a hidden state ht. The subscript t represents time in our notation. RNN's have a very elegant way of dealing with sequential (time) data that embodies connections between data points similar to the sequence [111]. Ref. [112] proposed recurrent neural network (RNNs) for sequential data processing such as voice and text processing. The defining characteristic of RNNs, which is distinct from that of RNN, the general neural networks are the introduction of the hidden state vector. The secret state represents the description of the previous input data which are modified once the new input is reached. Finally, after processing all input results, the secret state is the summarization of all sequences, which is similar to the processing of a sequence performed by a human being. Of course, RNN has the benefit of reading sentences that are read by a human. However, as the layer deepens, gradient explosions and vanishing problems occur, which can degrade performance [113]. Ref. [114] proposed the Long-Term Memory (LSTM) technique to avoid this. In order to prevent the gradient from bursting and causing disappearing problems, LSTM adds the cell state to change the previous knowledge. LSTM has been commonly

used for text classification because it can learn high-level representation using a deeper layer due to the cell status while maintaining the sequence of representations given by RNN. Ref. [115] applied LSTM to the emotion classification of short texts on social media. Ref. [116] suggested a densely connected Bi-LSTM composed of several Bi-LSTM layers, which shows improved efficiency than Bi-LSTM.

4. Deep neural networks (DNNs)

A neural network can be a deep neural network (DNN) with many layers that make it very mind-boggling. DNN contains one layer of data, at least one hidden layer, and one layer of output. A rectilinear unit (ReLU) is contained in a hidden sheet. ReLU is a mechanism for activation which has specified the positive part of its argument. ReLU has fewer gradient problems and is efficient in terms of computation. As each neuron in a single layer is connected with each neuron in the next layer, the secret layer is also called a fully linked layer [110]. A typical neural network (NN) consists of several neurons-called simple, interconnected network processors, each producing a series of activations of real value. Sensors that sense the environment activate input neurons, and weighted connections from previously active neurons trigger other neurons [117]. Ref. [118] presented a new tool for analyzing information obtained from X using deep neural networks to process cybersecurity.

5.3. Ensemble and Hybrid Learning

Ensemble learning is a powerful machine learning technique that improves model performance by combining the predictions of multiple individual models. The key principle is to leverage the diversity and strengths of these models to enhance prediction accuracy and robustness. Ensemble methods typically involve training several base models independently and then aggregating their predictions to arrive at a final output [119]. Ref. [120] developed a framework for identifying X bots using profile metadata. This study optimized the framework by comparing techniques for data preprocessing, feature selection, and model combination. The best results were achieved using Weight of Evidence encoding, Extra Trees for feature selection, and Random Forest blending, resulting in an impressive 93% AUC. While this approach offers rapid threat detection due to its reliance on static profile data, it is less effective than methods incorporating behavioral analysis. Ref. [121] developed a novel unsupervised ensemble learning method to detect previously unseen attacks in IoT networks using unlabeled data. The system generates labeled data for training a deep learning model to identify IoT attacks. Additionally, it employs feature selection to optimize attack detection. The proposed model effectively recognized attacks in unlabeled IoT data, with a Deep Belief Network (DBN) achieving a 97.5% detection accuracy and a 2.3% false alarm rate when trained on the generated labeled dataset. Ref. [122] conducted a study focused on detecting hate speech using machine learning and ensemble learning techniques during the COVID-19 pandemic. The research utilized X data, which was extracted via the platform's API with the aid of trending hashtags relevant to the pandemic. To facilitate analysis, tweets were manually annotated into two distinct categories based on various factors. Feature extraction was performed using methods such as TF-IDF, Bag of Words, and tweet length. The study identified the Decision Tree classifier as particularly effective, achieving precision of 98%, recall of 97%, an F1-score of 97%, and an accuracy of 97%. However, the Stochastic Gradient Boosting classifier demonstrated superior performance overall, with a precision of 99%, recall of 97%, an F1-score of 98%, and an accuracy of 98.04%. Ref. [123] explored the potential of deep learning for detecting novel cyber threats—those unseen during model training. The study also examined the role of bias in identifying these unknown attacks. Traditional machine learning models, limited by single datasets, often struggle with unforeseen threats,

exhibiting high accuracy in familiar scenarios but failing to recognize the unfamiliar. To address this, the research proposed a more adaptable Intrusion Detection System (IDS) using an ensemble of deep learning classifiers. Trained on multiple benchmark datasets, this ensemble aimed to detect unknown attacks without prior knowledge of specific threat patterns. By combining proven classifiers for sequential data, the research sought to create a robust IDS capable of identifying a wide range of cyber threats. The results demonstrated the effectiveness of this approach, offering promising performance and advancing practical IDS solutions. Ref. [124] developed a novel ensemble stacking learning approach to detect cyberbullying on X. The method integrates multiple deep neural networks (DNNs) and introduces a modified BERT model, BERT-M. The study employed a preprocessed X dataset and utilized word2vec embeddings generated by Continuous Bag of Words (CBOW) to extract features. Convolutional and pooling layers processed these features to capture offensive language patterns. The proposed stacked model and BERT-M achieved exceptional performance, surpassing existing NLP cyberbullying detectors. With an F1-score of 0.964, precision of 0.950, and recall of 0.92, the stacked model demonstrated high accuracy in detecting cyberbullying within 3 min. The ensemble approach yielded a detection accuracy of 97.4% on the X dataset and 90.97% on a combined X and Facebook dataset, emphasizing its effectiveness in combating cyberbullying across platforms. Ref. [125] employed an ensemble approach to accurately classify crime-related tweets. Data were collected using the Tweepy and Twint libraries and processed with TF-IDF vectorization. The ensemble combined Logistic Regression, Support Vector Machine, k-Nearest Neighbors, Decision Tree, and Random Forest classifiers (weighted 1, 2, 1, 1, and 1, respectively) using a soft weighted Voting classifier. This methodology achieved an impressive 96.2% accuracy on the test dataset, demonstrating the effectiveness of the ensemble for crime tweet classification. Ref. [126] identify and classify spam URLs on X developed multiple models using a combination of URL content, user profile information, and hybrid features. A large X dataset was analyzed to create comprehensive feature sets for training various ensemble learning models. Our models achieved high accuracy, often exceeding 90%, particularly when using k-Nearest Neighbors within bagging and Random Forest ensembles. Results indicate that combining user profile, content, and hybrid data significantly enhances spam detection accuracy. Ref. [127] research delves into real-time public opinion by analyzing tweets across a wide range of topics, including COVID-19, crime, spam, Flipkart, migraine, and airlines. The study harnessed the X API to collect a substantial dataset of tweets, which were then meticulously cleaned and preprocessed using natural language processing (NLP) techniques. To gauge public sentiment, a comparative analysis was conducted using both traditional machine learning (ML) algorithms (Naïve Bayes, Decision Trees, Random Forest, Logistic Regression) and advanced deep learning (DL) models (recurrent neural networks, Long Short-Term Memory, Gated Recurrent Units). While these models were evaluated independently, the core contribution of the research lies in a novel ensemble approach that combines ML and DL models. Ref. [128] focused on automating the detection of binary labels in aggressive tweets, a novel system has been developed, demonstrating exceptional performance relative to previous studies conducted on the same dataset. The study employed a stacking ensemble machine learning approach, integrating four distinct feature extraction techniques to enhance performance within this framework. By combining five machine learning algorithms—Decision Trees, Random Forest, Linear Support Vector Classification, Logistic Regression, and k-Nearest Neighbors—into an ensemble model, the researchers were able to achieve significantly improved results over traditional machine learning classifiers. The stacking classifier attained an impressive accuracy rate of 94.00%, which not only surpassed the performance of conventional models but also outperformed the results of earlier experiments using the identical dataset. The findings

highlighted the system's effectiveness, achieving an accuracy rate of 94.00% in correctly classifying tweets as either aggressive or non-aggressive. Ref. [129] developed a sophisticated deep learning model tailored for cyberbullying detection in tweets. Leveraging the label X_parsed_dataset.csv, the model extracted keywords and entities using Maximum Entropy. A 1D-CNN architecture was then applied to classify tweets as truculent or non-truculent. The study compared four preprocessing methods (Unigram, Bigram, Trigram, and N-gram) and achieved impressive results: 96.1% accuracy, 93.6% precision, 73.7% recall, and an F1-score of 83.8% across different evaluations. Ref. [130] research introduces a novel cybersecurity approach, IRSO-EDLCS, to bolster cyber attack detection in Industrial Internet of Things (IIoT) environments. This technique leverages an Improved Reptile Search Optimization (IRSO) algorithm for feature selection, optimizing feature relevance for enhanced detection accuracy. An ensemble of Deep Belief Network (DBN), Bidirectional Gated Recurrent Unit (BiGRU), and Autoencoder (AE) models is then employed to identify cyber threats. To further refine the model, a Modified Gray Wolf Optimizer (MGWO) is integrated for hyperparameter tuning, maximizing the ensemble's performance. Rigorous simulations on a benchmark database demonstrate IRSO-EDLCS's superior performance compared to existing methods, highlighting its potential to significantly advance IIoT cybersecurity, Table 1 presents the summary of the related work.

Table 1. A summary of related work.

Authors	Year	Focus Area	Techniques/Models	Results
[120]	2021	X bot detection	Weight of Evidence encoding, Extra Trees (feature selection), Random Forest (blending)	93% AUC; rapid threat detection with static profile data but less effective than behavioral analysis methods
[121]	2022	IoT network attack detection	Unsupervised ensemble learning, Deep Belief Network (DBN)	97.5% detection accuracy, 2.3% false alarm rate
[122]	2022	Hate speech detection during COVID-19	Decision Tree, Stochastic Gradient Boosting, TF-IDF, Bag of Words, tweet length	Stochastic Gradient Boosting: 99% precision, 97% recall, 98% F1-score, 98.04% accuracy
[123]	2022	Cyber threat detection	Ensemble of deep learning classifiers	Effective detection of novel cyber threats, adaptable Intrusion Detection System (IDS)
[124]	2023	Cyberbullying detection on X	Ensemble stacking, deep neural networks (DNNs), BERT-M, word2vec, CBOW	97.4% accuracy on X dataset, 90.97% on combined X andFacebook dataset, F1-score: 0.964, precision: 0.950, recall: 0.92
[125]	2023	Crime-related tweet classification	Logistic Regression, Support Vector Machine, k-Nearest Neighbors, Decision Tree, Random Forest, TF-IDF	96.2% accuracy using soft weighted Voting classifier
[126]	2024	Spam URL detection on X	k-Nearest Neighbors, bagging, Random Forest, URL content, user profile, hybrid features	High accuracy (>90%) when using combined feature sets

Table 1. *Cont.*

Authors	Year	Focus Area	Techniques/Models	Results
[127]	2024	Real-time public opinion analysis	Naïve Bayes, Decision Trees, Random Forest, Logistic Regression, RNN, LSTM, GRU, ensemble of ML and DL models	Comparative analysis of ML and DL models; novel ensemble approach combining ML and DL
[128]	2024	Aggressive tweet detection	Stacking ensemble, Decision Trees, Random Forest, Linear SVC, Logistic Regression, k-Nearest Neighbors	94.00% accuracy in classifying tweets as aggressive or non-aggressive
[129]	2024	Cyberbullying detection in tweets	1D-CNN, Maximum Entropy, Unigram, Bigram, Trigram, N-gram	96.1% accuracy, 93.6% precision, 73.7% recall, 83.8% F1-score
[130]	2024	Cyber attack detection in IIoT environments	IRSO algorithm, Deep Belief Network (DBN), BiGRU, Autoencoder (AE), Modified Gray Wolf Optimizer (MGWO)	Superior performance in IIoT cybersecurity, effective feature selection and hyperparameter tuning

6. X Security: ML/DL Solutions

In this section, the focus is on identifying key vulnerability characteristics and conducting a comprehensive literature review of prior research studies that have utilized DL and X data for detecting cyber attacks. After providing a brief overview of vulnerability detection and exploitation, we will delve into a detailed examination of these previous studies.

6.1. Detection of Vulnerabilities and Exploits on X

Vulnerabilities and exploits are problematic security weaknesses. Vulnerabilities are typically found within software systems, while exploits arise because of these vulnerabilities. In other words, exploits are the actual manifestation of the vulnerabilities within software systems. To better understand these weaknesses, further research is necessary to identify the root cause of these security issues and develop effective mitigation strategies [131]. To prioritize the protection of systems, this section examines the marginal variance between various weaknesses concepts. The scope of this investigation is centered on utilizing X data as a source of information to identify any new vulnerabilities or to assess the presence of exploits targeting known vulnerabilities. Given that most security breaches are subject to temporal constraints, it is imperative to have effective mechanisms in place for detecting such incidents in a timely manner. By doing so, it becomes possible to prioritize the allocation of resources towards rectifying the vulnerability, thereby saving valuable time and effort [132]. Researchers utilize common vulnerability and exposure identifier (CVE-ID) as a feature to predict the likelihood of exploitation for known vulnerabilities. A novel approach was presented by [133] for the generation of early warnings for real-world exploits against known vulnerabilities. The prediction is grounded in an analysis of tweets that mention these vulnerabilities, along with their associated CVE-IDs, in the context of malicious intent. To achieve this, they utilized X's Streaming API to monitor occurrences of the keyword "CVE". Additionally, they employed the SVM algorithm, a supervised machine learning technique, to develop a classifier that leverages user and tweet-related features to identify emerging cyber attacks. The results of this approach demonstrated superiority over the commonly recommended vulnerability scoring system (CVSS), with a reduced rate of false positives. Furthermore, the method was capable of

detecting exploits with a median lead time of two days ahead of existing datasets. Ref. [134] proposed a method of utilizing social media analysis for software vulnerability monitoring in the HANA (SMASH) architecture. The SMASH process involves conducting a search for security and vulnerability terminologies as well as software components from sources such as X and the National Vulnerability Database (NVD) and storing the information in a local database. Subsequently, tweets are grouped together through the utilization of a modified K-mean clustering algorithm, which takes into account the context of each tweet. The NVD serves as a reference point to differentiate between old and new information regarding vulnerabilities. Currently, this process is conducted manually. The results of this research showed that 100% of the NVD weaknesses were mentioned on X, with 41% of Common Vulnerabilities and Exposures (CVEs) being published on X prior to the official NVD release, with an average of 20 days in advance. Additionally, approximately 75% of Linux–Kernel zero-day vulnerabilities were disclosed on X before the official disclosure, with an average lead time of 19 days. Ref. [135] proposed a novel crowd-sourced vulnerability detection system that utilizes X as the main source of real-time information. The system employs the use of security-specific keywords to identify tweets that pertain to potential security incidents or anomalies in online services or accounts. Subsequently, the proposed model compares these tweets with the vulnerability descriptions present in the Common Vulnerabilities and Exposures database (CVE-DB) to determine whether the detected behavior constitutes a new vulnerability or a zero-day exploitation of a previously known vulnerability. Ref. [136] utilized a corpus of tweets posted by security experts to construct a Support Vector Machine (SVM) classifier with the objective of segregating tweets that contained security alerts and software patch/fix information from general security discussions. The classifier was developed utilizing three sets of word frequency features: unigram, bigram, and a combination of both. The study found that the proposed model had an accuracy rate of 94% when classifying tweets over a one-year time period [137]. However, the authors noted that the methodology, which is based solely on word appearance in tweets, can result in misclassification of tweets as false positive. This occurs when security-related words appear in both security-related and non-security-related tweet phrases, leading to the misclassification of general discussions as useful alerts, and vice versa. Ref. [137] introduced a cascaded convolutional neural network (CNN) framework for identifying and categorizing cyber attack-related events on X. This approach involves two CNN models: a binary classifier to distinguish cyber-related from irrelevant tweets, followed by a multi-class classifier to assign specific threat labels (DDoS, zero-day vulnerabilities, ransomware, data leaks, or marketing/general) to the identified cyber tweets. The model was trained on a dataset of approximately 21,000 annotated tweets. The model achieved an average F1-score of 0.82 in classifying cyber threats. Ref. [138] developed a Random Forest model to automatically classify cyber threats using X data, achieving an accuracy of 80%. Ref. [139], the authors addressed the discrepancy between the CVE-DB and the findings of [134] by developing a method for identifying security-related tweets that contain information about vulnerabilities, even if the specific vulnerability ID is not mentioned. To do this, they propose a model that leverages the CVE-DB to learn the features of vulnerabilities through the use of a centroid classifier. The model is trained using descriptions of vulnerabilities as positive samples. The pipeline begins by collecting tweets from specialized security accounts and extracting TF-IDF features for each tweet. The tweets are then passed through the trained model and classified as normal or not based on their distance from the centroid and a specified threshold value. The performance of the model was evaluated using a manually labeled dataset, yielding an F1-score of 64%, surpassing the results of SVM, MLP, and CNN baseline models. Ref. [140] developed machine learning models to categorize cybersecurity-related X accounts. They collected cybersecurity-related tweets

using X's Sampling API and manually labeled them. A baseline model was trained to identify general cybersecurity accounts, followed by sub-models for classifying accounts into individuals, hackers, or academia. Four machine learning models (Decision Tree, Random Forest, Support Vector Machine, and Logistic Regression) were compared using various account features. Random Forest achieved the highest performance, with 93% accuracy for the baseline model and 88–91% accuracy for the sub-models. Ref. [141] introduced Darkintellect, a machine learning-based approach for identifying cyber attacks through X data, The researchers collected approximately 21,000 cyber attack-related tweets using the Tweepy3 Python library. To prepare the data for analysis, they employed NLP techniques to preprocess the tweets by removing irrelevant information and special characters. Feature extraction was performed using TF-IDF to represent the text data numerically. Five machine learning algorithms—SVM, RF, DT, XGBoost, and AdaBoost—were evaluated for their ability to classify tweets as cyber attacks or not. The results demonstrated that Decision Trees (DT) outperformed the other methods, achieving a classification accuracy of 87.54%. The authors of [142] developed a hybrid NLP and CNN model to identify and categorize four cyber attack types (malware, phishing, spam, and bot attacks) within social network messages. This method uniquely focuses on textual analysis, making it adaptable across different platforms. Evaluated on real-world data, the model underwent a two-phase process. First, it detected the presence of any cyber attack, followed by classification into a specific category. The model achieved an overall accuracy of 82%. The efficiency of utilizing information disclosed on X to detect zero-day vulnerabilities and exploits has been established through several studies. However, it is noted that these solutions have limitations in terms of vulnerabilities, where a more comprehensive approach to detecting security content is necessary. This is since the retrieved tweets may not include specific vulnerability numbers, unlike the advanced counting-based secret-sharing security technique.

6.2. Detection of Security Content

X is a popular social media platform used by millions of users worldwide to share information. However, there is a concern about the spread of false information or propaganda by malicious actors with security implications. To address this, researchers have proposed various methods for detecting security-related content on X, including natural language processing, ML algorithms, and human expert judgment. These proposals aim to provide a means for detecting malicious content and ensuring the security and reliability of information disseminated on X. In this section, we will examine a body of literature that has put forth proposals aimed at detecting security-related content that is disseminated on X. In their study [143]. Introduced a novel framework for detecting localized events by analyzing the presence of bursty keywords and the spatial distribution of documents. The authors emphasized the importance of incorporating both temporal and spatial aspects in event detection. The framework they presented is based on the assumption that the occurrence of bursty keywords and their spatial distribution can provide useful clues in identifying localized events. This framework is designed to identify localized events in real time and can provide valuable information to decision makers in various domains. The authors' innovative approach to event detection has the potential to revolutionize the way we analyze and understand events, providing a more comprehensive and accurate picture of what is happening on the ground. The Cyber X framework by [144] is a profile-based system that utilizes X's API to retrieve tweets that pertain to security-related topics and system profiles. The Security Vulnerability Concept Extractor (SVCE) is then employed to extract terms that are related to security vulnerabilities, with only tweets that contain two or more terms being retained for further analysis. These terms are tagged with technical descriptors, such as means of attack, consequences, affected software and hardware, and

version numbers. The output from the SVCE is mapped to real-world concepts using DBpedia and YAGO ontologies, as well as the Unified Cybersecurity Ontology (UCO) to provide context-specific knowledge. The resulting data are stored as triples in a local Cybersecurity Knowledge Base (KB) and a set of Semantic Web Rule Language (SWRL) rules are added to the system to reason over the KB and generate alerts based on potential threats and vulnerabilities. As shown in Figure 4, the rules interpret the relationships between elements in the KB and update the system's state when new tweets arrive, triggering alerts when threats are detected.

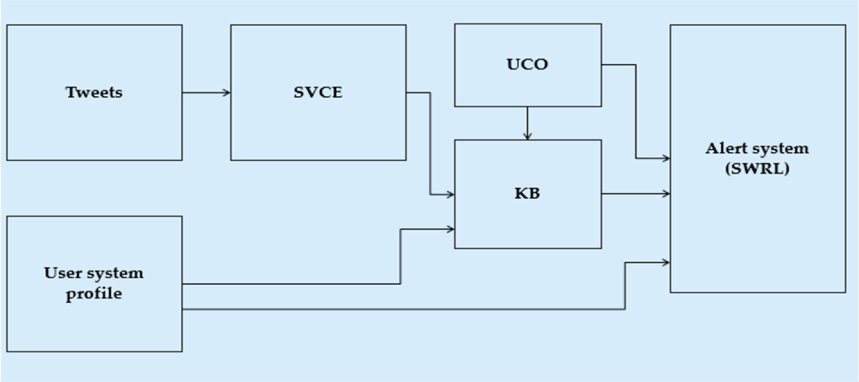

Figure 4. Cyber-Twitter architecture, adapted from [145].

These alerts are reviewed by cybersecurity experts and used to inform their security policies as necessary. In a ten-day experimental evaluation, it was found that 13 out of 15 alerts were considered useful by assessors. However, the framework is limited in its ability to detect the context of tweets, which is crucial in differentiating tweets that discuss emerging attacks from those that discuss security-related topics in general. In the study by [146], a real-time text-mining approach was utilized to detect unfamiliar security terms from tweets posted by a predetermined set of 69 security experts. The process involved retrieving tweets every 60 min, filtering unique words and excluding words that appeared in dictionaries of English, stop-words, technical terms, and Italian terms. The remaining words were considered new security threats if they were mentioned in the same tweet with terms contained in a threat dictionary. The generated information included the new term, its volume of mentions in the past 60 min, the contents of the posts, and related words. The entire process from retrieving tweets to generating warnings took approximately 0.6 s, resulting in an accuracy rate of 84%. One notable example was the detection of the Mirai term 49 days prior to the actual attack in October 2016. However, the solution is limited in its ability to detect new attack terms, as it relies on the knowledge of the experts who may not be aware of evolving security threats until they become public, or the attack does not have an unfamiliar name prior to occurrence. Figure 5 in the study provides a visual representation of the proposed algorithm. Ref. [147] introduced a real-time security event detection system that utilizes a taxonomy of cybersecurity events and corresponding seed keywords to identify security-related tweets. Over a nine-month period, the system collected 47.8 million tweets utilizing seed keywords ranging from unigrams to 6-g. To further refine the results, a blacklist of phrases was added to reduce false positives. SONAR groups similar tweets into clusters using cosine similarity and locates the geographic area of high discussion through the use of Google Map Geocoding API. Additionally, it includes a keyword finder that continuously updates the keyword list to remain relevant. The system utilizes GloVe embedding to find semantically related words, allowing for scalability while still relying on the analyst's final decision. The efficiency of SONAR was evaluated with

positive results, showing that approximately 25% of the detected security events were relevant. The architecture of SONAR is presented in Figure 6.

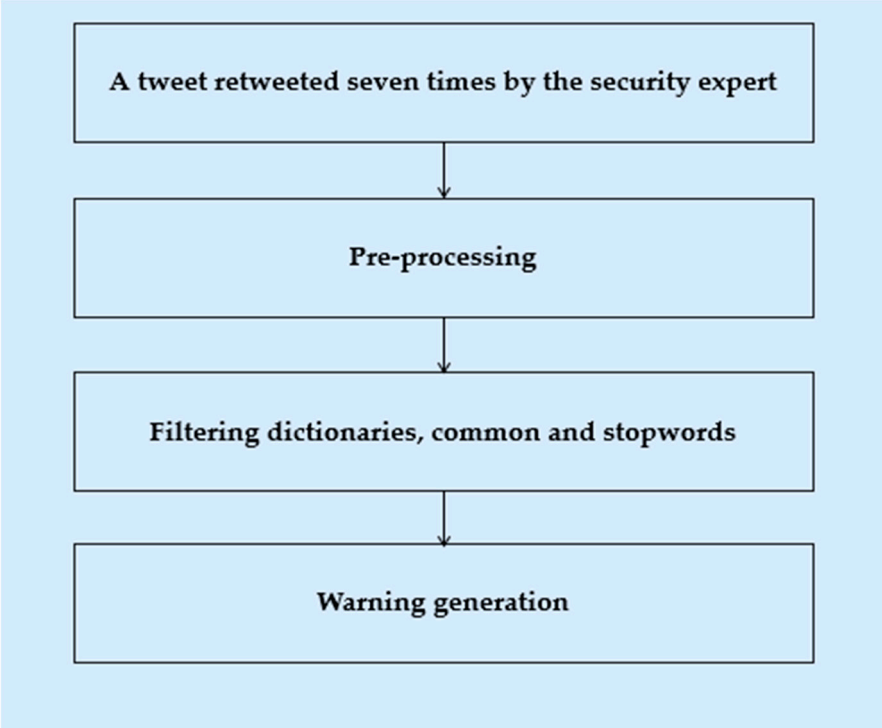

Figure 5. Running example, adapted from [146].

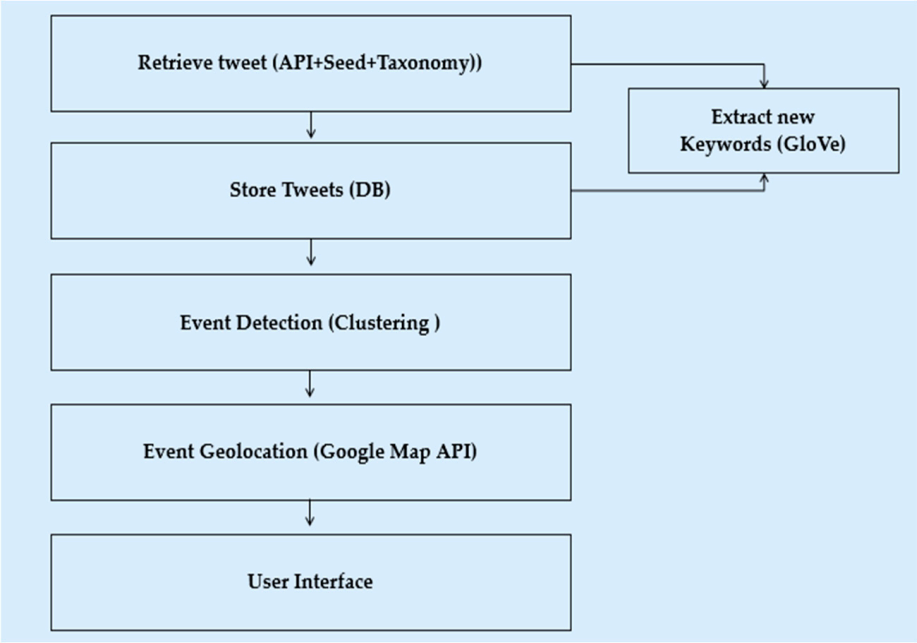

Figure 6. SONAR architecture, adapted from [147].

The SYNAPSE system [148] is a real-time security event detection tool for IT infrastructure. SYNAPSE utilizes a dataset of over 195,000 tweets, retrieved from two sets of X accounts publishing security-related tweets, designated as S1 and S2. The tweets from the S1 accounts form the training set, while the validation and testing sets are comprised of tweets from both S1 and S2 X accounts, allowing for the possibility of adding more accounts in the future. The tweets are filtered based on security keywords representing

three different infrastructures and are then processed using TF-IDF features, a supervised ML approach using MLP and SVM algorithms, and a dynamic stream clustering methodology to group similar tweets into events. The final output of the model is presented in the form of an indicator of compromise (IoC) for manual inspection or integration with threat intelligence tools. The evaluation results showed that SVM achieved a better balance between true positive and true negative classification rates, approximately 90%, and the IoC was evaluated for relevance and modernity, demonstrating the efficiency of the end-to-end model. Figure 7 illustrates the main steps of the SYNAPSE system.

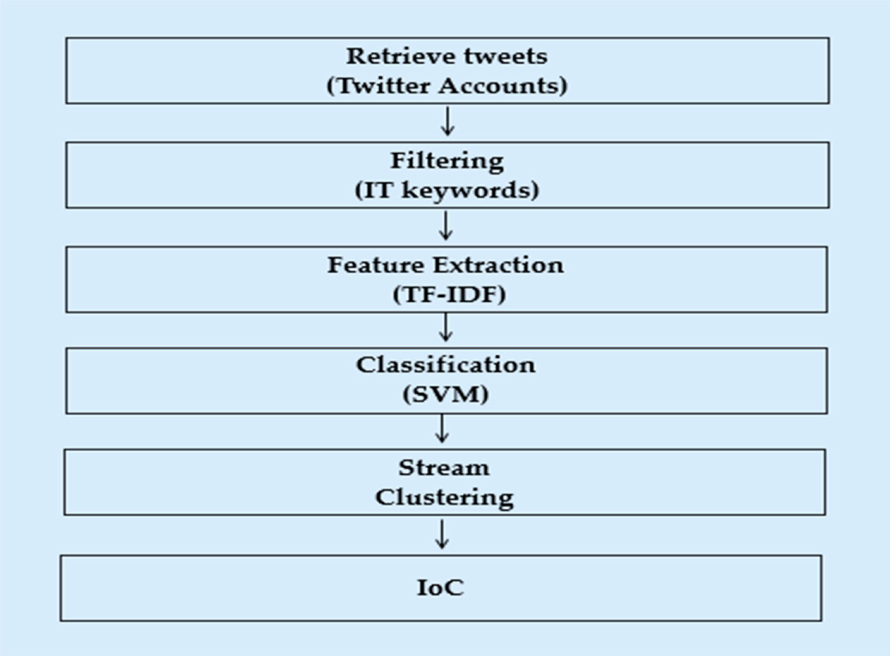

Figure 7. SYNAPSE's architecture, adapted from [148].

In the study conducted by titled DataFreq, the authors present a novel approach for tracking the sentiment score of a particular company in relation to the probability of a potential security breach as shown in Figure 8. To gather relevant information, 70,475 tweets from a set of security expert accounts were collected and filtered based on a security keyword list. A novel ML model was developed using the Logistic Regression algorithm with n-gram feature extraction technique to classify the sentiment of the tweets. The authors also aimed to update the keyword list regularly to ensure that the system can adapt to new security-related terms. Through their experiments, the model successfully identified three new words that indicated potential security breaches. The model's performance was evaluated over a four-week period, during which an increase in phishing attacks was observed during the third week (a holiday). The results indicate the effectiveness of the proposed method, with an 85% precision, 84% recall, and F1-score, reflecting the real-world scenario of an increased number of phishing attacks during holidays. The results are presented to the end user in an easily interpretable format in real time, allowing security analysts to easily monitor the average sentiment score of the company and the most frequently mentioned security issues. Ref. [149] combined the moving threshold average algorithm with the Gaussian tweet sentiment signal detection algorithm and the top hashtag. Analysis algorithm to develop a new sentiment analysis model for X data. The proposed model was able to effectively identify the sentiment of tweets, and the hashtags associated with them. The results showed that the proposed model outperformed traditional sentiment analysis algorithms in terms of accuracy and efficiency. The study also showed that the combination of the moving threshold average algorithm and the Gaussian

tweet sentiment signal detection algorithm improved the accuracy of sentiment analysis by detecting the sentiment signal more effectively. The top hashtag analysis algorithm helped to identify the hashtags associated with the sentiment and provide a more comprehensive analysis of the sentiment expressed in the tweets. This study highlights the importance of combining multiple algorithms to improve the accuracy of sentiment analysis in social media data.

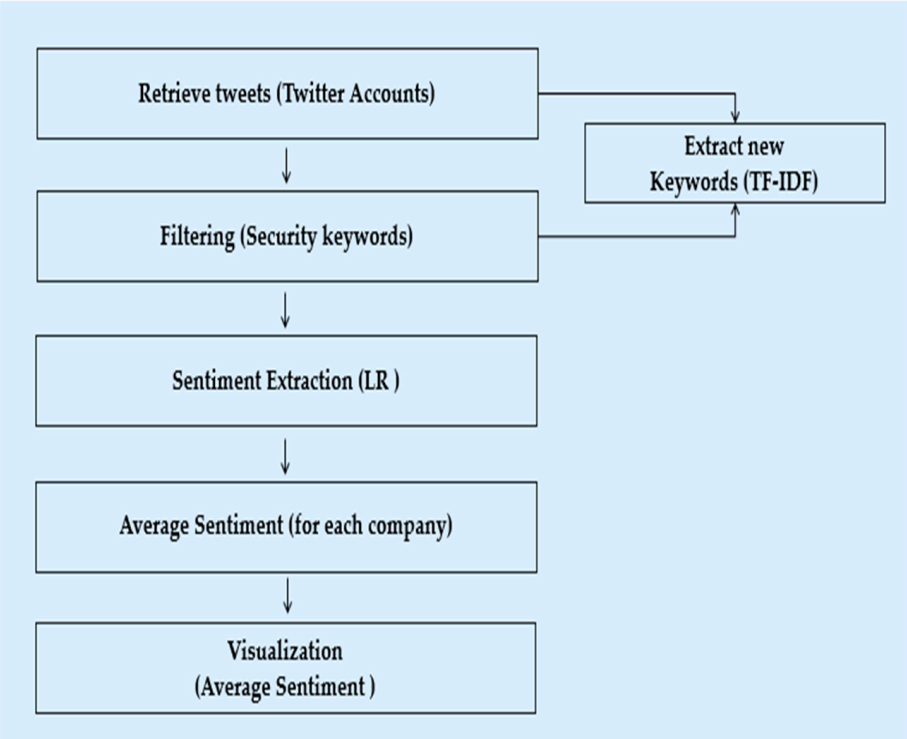

Figure 8. Data mining situational awareness scheme, adapted from [150].

Ref. [151] presented a X-based framework for detecting traffic incidents as a supplementary method for monitoring traffic conditions. This framework involved the collection of a large volume of tweets using X API endpoints, which were then labeled through a systematic and efficient process that incorporated shortcuts for speed without sacrificing data quality. The labeled data were used to develop a DL model for detecting traffic-related events from X streams. The tweets were transformed into numerical feature matrixes using word-embedding models, and CNN and RNN architectures were utilized to distinguish traffic-related tweets. The results of the experiments showed that the proposed model outperformed existing models in the field. However, to fully implement the framework, a geocoder must be developed to identify the location of traffic events and disseminate the relevant information to users in real-time. This system can provide benefits to travelers by helping them choose the most efficient routes, as well as assisting traffic management agencies in restoring smooth traffic flow by detecting unexpected changes in traffic flow characteristics. The study by [118] investigates the real-time detection of security information relevant to IT infrastructure, as shown in Figure 9. In this work, tweets from two sets of security-related accounts are collected and processed in three-time intervals. The tweets are filtered based on keywords and transformed into numerical representations using word2vec word embedding. These embedded tweets are then input into three parallel CNN layers for classification. The output of this model is a binary classification of each tweet as security-related or not, followed by Named Entity Recognition (NER) to extract key entities such as company, asset, vulnerability, or IDs using a bidirectional Long Short-Term Memory Network (BiLSTM). The extracted entities are then utilized to generate

Indicator of Compromise (IoC) alerts. The classification performance was evaluated and achieved a true positive rate (TPR) of 94% and a true negative rate (TNR) of 91%, while the NER achieved a F1-score of 92% in specifying the correct labels. A comparison between the discovered IoCs and traditional security databases such as the National Vulnerability Database (NVD) revealed that the model was able to detect vulnerability information 1 to 149 days ahead of NVD.

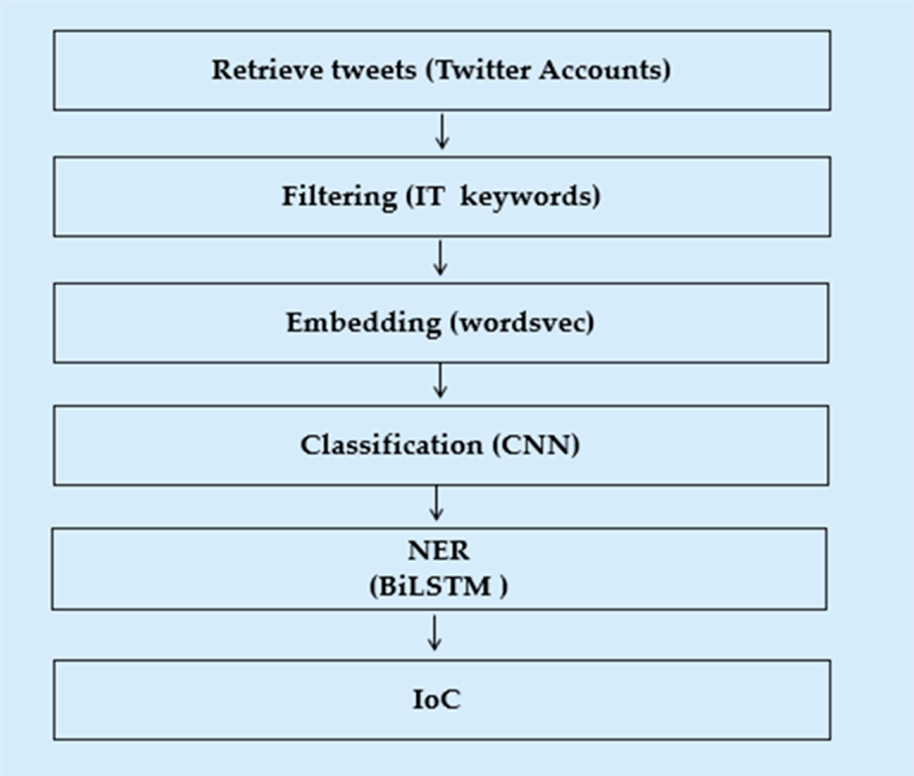

Figure 9. DeepNN BiLSTM architecture, adapted from [118].

Ref. [14] proposed a multitask learning approach based on the iterated dilated convolutional neural network (IDCNN) and Bidirectional Long Short-Term Memory (BiLSTM) natural language processing technology. The ML algorithm was presented to set up a highly accurate network model. Ref. [152] used locality-sensitive hashing to roughly find related items and incremental clustering to implement a realistic, real-time event detection algorithm. Researchers are trying to define the features of tweets and use suitable algorithms to solve the problems they are researching. Ref. [153] show a framework for classifying OSINT data into cybersecurity-related to be introduced and analyzed, and the accuracy of those data was subsequently improved using an unsupervised method. Ref. [154] conducted a study where they utilized supervised classification methods to identify spammers on X. Data were gathered from Tweepy API and consisted of 2798 accounts in the training group and 578 accounts in the testing group. Eighteen features were extracted from user profiles. Extreme Machine Learning (EML) showed the highest accuracy at 87.5%. Ref. [155] utilized ML to identify fraudulent X accounts. They employed several ML algorithms, including SVM, LR, RF, and KNN, and considered six account metadata features: likes, language code, sex code, status count, friends count, followers count, and favorites count. To enhance the accuracy of these algorithms, they applied two different normalization techniques: Z-score and Min–Max. Through their method, they achieved impressive accuracy levels of 98% for both the RF and KNN models. Ref. [156] suggested three effective techniques for identifying fraudulent accounts in their study. The classification algorithms

employed included Linear and Radial SVM, RF, and KNN. The dataset utilized contained 3964 entries. RF yielded more reliable predictions by resolving the overfitting issue. The RF's K-Fold Cross-Validation Scores showed a mean of 0.979812 and a standard deviation of 0.019682. By contrast, Radial SVM was unsuccessful and resulted in more false negatives. Nonetheless, using the Ensemble approach resulted in better accuracy. The study conducted by [157] aimed to distinguish live tweets as spam or ham and conduct a sentiment analysis on both live and stored tweets, categorizing them as positive, negative, or neutral. The methodology involved utilizing two datasets from Kaggle and extracting sentiment features using vectorizers like TF-IDF and BoW models. The extracted features were then input into various ML and DL classifiers. The best performing classifiers were LSTM in both spam detection with a 98.74% accuracy rate and sentiment analysis with a 73.81% accuracy rate. In a recent study by [158], a method for identifying X cyber threats was introduced. Their dataset consisted of 15.6 million tweets, with 3.2 million accounts sent during the US Elections, and they used the XGBoost algorithm to select 229 features from approximately 337 user-extracted features. The researchers trained and validated three ML models, including SVM, RF, and XGBoost, and found that XGBoost had the best performance. Their findings suggest that XGBoost is better at generalizing to new data than the other models. The study reported only a small decrease in performance, with a 2% decrease in F1-score from 0.916 to 0.896 and a 0.03% decrease in ROC-AUC from 0.98 to 0.977. Ref. [159] Identified OSN threats and recommended a double-factor authentication method using digital face-processing after entering the password using Matlab. They achieved a 95% accuracy rate by applying DL classification to a real dataset from the live webcam to train the model. They also mentioned the problem of fake accounts and its significant impact on executing spam campaigns and spreading malware and phishing attacks. The authors conducted a study on detecting fake and legitimate profiles on OSN, using two datasets from Facebook and Instagram. They applied ML algorithms such as Naive Bayes, Logistic Regression, Support Vector Machines, k-Nearest Neighbor, Boosted Tree, neural networks, SVM Kernel, and Logistic Regression Kernel, and found that SVM achieved the highest classification accuracy for the Fake Profiles detection datasets with 97.1%. Table 2 summarizes the comparisons of previous studies on the detection of cybersecurity threats in X. All previous studies indicate the focus on the usage of effective classifiers to improve detection accuracy.

Table 2. Comparison of previous studies on the detection of cybersecurity threats on X.

Study	Focus	Methodology	Datasets
[160]	- Events from X that requires only minimal supervision - DoS attacks, data breaches, and account hijacking	Weakly supervised learning	Tweets containing "DDoS"
[147]	An automatic, self-learned framework that can detect, geolocate, and categorize cybersecurity events in near-real time over the X stream	First story detection	Streaming tweets
[161]	Machine learning techniques by considering user behavior, content of tweets, social relationships, etc., to detect different types of cyberthreats	SocialKB	- Tweets containing "URLs" - Streaming tweets

Table 2. *Cont.*

Study	Focus	Methodology	Datasets
[137]	Cybersecurity events	Deep learning model; cascaded CNN architecture	Labelled 21,000 tweets collected using Tweepy
[162]	A novel application of NLP models to detect denial of service attacks using only social media as evidence	Basic neural network	Tweets written on attack day
[163]	Treat the event detection problem in a multimodal X hashtag network	Expectation-maximization (EM) algorithm	Tweets containing hashtag
[118]	A novel tool that uses deep neural networks to process cybersecurity information received from X	SVM, MLP, CNN, BiLSTM	Tweets filtered by keywords
[164]	Analyze the severity of cybersecurity threats based on the language that is used to describe them online	Supervised ML models	Tweets containing "DDoS" and "vulnerability"
[165]	Cybersecurity-related data	Three supervised ML models; SVM, MNB, and RF	Real-time cyber attack Data from HuffPost News Site
[138]	Cybersecurity threats relevant data	RF	Filtered tweets collected using X's streaming API
[139]	Collection method of Cyber threat tweets	Centroid, One-class SVM, CNN, LSTM	Streaming Tweets
[16]	A multitask learning approach combining two Natural Language Processing tasks for cyberthreat intelligence	Multitask Learning (MTL)	Streaming Tweets
[113]	A novel word embedding model, called contrastive word embedding, that enables to maximize the difference between base embedding models	CNN, RNN and LSTM	Curated data, OSINT data, and background knowledge
[14]	- Detection of cyber threat events on tweets. - Named Entity Recognition (NER) for tweets	Multitask learning NLP, IDCNN, BiLSTM	Streaming Tweets
[140]	Cybersecurity-related discussions	Four supervised machine learning models; Decision Tree, Random Forests, SVM, and Logistic Regression	Labelled tweets collected using the X Sampling API
[141]	Cybersecurity threats relevant data	Five ML models: SVM, Random Forest, Decision Tree, XGBoost and AdaBoost	Labelled 21,000 tweets were collected using a python package Tweepy
[142]	Cybersecurity-related data	Deep learning model; CNN architecture	Social network messages

7. Analysis of X Cyber Threat Solutions

Performance metrics such as the confusion matrix, accuracy, recall, precision, F1-score, and PR curve are utilized to evaluate the effectiveness of various deep learning (DL) algorithms. The confusion matrix serves as a tool to represent different metrics, balancing

selectivity and specificity, with the goal of minimizing the time spent on Type I and Type II errors. The output of machine learning (ML) and DL algorithms is assessed through the confusion matrix, which helps assign input data to distinct labels. Widely regarded for its simplicity and effectiveness, the confusion matrix compares expected and observed values, using four key elements: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

In this context, a true positive (TP) occurs when both observed and expected values are positive, such as in the detection of cyber threats. A true negative (TN) occurs when both observed and predicted values are negative, indicating the absence of cyber threats. A false positive (FP) occurs when the observed value is negative, but the expected value is positive. Conversely, a false negative (FN) is when the observed value is positive, but the predicted value is negative.

Accuracy, which is the inverse of the error rate, is a commonly used performance metric in classification tasks. However, it may not be ideal for imbalanced datasets. Recall, on the other hand, measures the ability of a classifier to correctly identify true instances of a cyber threat. Recall is defined as the proportion of actual positives that the model successfully identifies and is positively correlated with the number of correct predictions from the minority class. Precision, meanwhile, measures the effectiveness of the classifier in predicting positive outcomes when it expects a positive result.

The F1-score combines both precision and recall into a single metric, representing their harmonic meaning. This metric is particularly useful for evaluating the performance of models on minority classes and is widely used in information retrieval systems.

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is another important metric, especially in cases of unbalanced data. The AUC-ROC allows for the assessment of model performance by considering the trade-off between true positive rates and false positive rates. The F1-score can also be employed to evaluate the training examples by integrating both precision and recall. Additionally, a curve illustrating the relationship between recall and false positive rate can be plotted, with better model performance correlating with a higher AUC value. Further details on calculating and interpreting PR curves are provided by [166]. In this study, precision, recall, and F1-score were used to assess classifier performance, with the specific formulas shown in Table 3.

Table 3. Performance measures.

Metrics	Description	Equation	Range
Accuracy (A)	Assess the number of TPs	$A = \frac{TN+TP}{TN+FN+TP+FP}$	[0–1]
Recall	The ratio of TP to a TP and FN	$R = \frac{TP}{TP+FN}$	[0–1]
Precision	The ratio of TP to a TP and FP	$P = \frac{TP}{TP+FP}$	[0–1]
F1-Score	Combines precision and recall	$FI = 2\frac{P-R}{P+R}$	[0–1]
AUC	The area between two points bounded by the function and the x-axis	$AUC = \int_a^b f(x)dx$	[0–1]

In this section, a detailed comparison is presented for the studies previously introduced in the section on security content detection: SYNAPSE, DeepNN, DataFreq, CyberX [144], text mining, and SONAR [147]. The reason why a detailed comparison is presented for six studies previously introduced is to provide a comprehensive analysis of the similarities and differences between the studies. This comparison may help to identify patterns, trends, and gaps in the existing research, as well as to determine which approaches or methodologies have been most effective in addressing the research questions. Presenting a detailed comparison of the six studies is a valuable approach to gaining a deeper understanding

of the existing research and informing future research directions. These studies were chosen to meet the objectives of analyzing security content on X without relying on specific vulnerabilities. Various prediction factors were compared, and their measurements were combined to provide an overall assessment and recommendation for real-time usage. The study also provides insight for future research aimed at improving models to increase the probability of cybersecurity success. The prediction factors assessed included scope of detecting the investigation of semantic characteristics, the complexity of the algorithm, degree of information condensation, scalability, and effectiveness. The results for each factor were summarized for all methods, revealing interesting comparison figures. The evaluation of the factors was further analyzed to provide the pros and cons of each model for further investigation and combination.

7.1. The Complexity of the Algorithm

One perspective for comparison involves the use of classification algorithms. The most basic technique involves filtering tweets based on specific keywords. The text mining study by [146] filters tweets using technical, security, and English dictionaries to extract unfamiliar terms. In CyberX [144], a Semantic Web Rule Language (SWRL) is utilized to issue alerts related to the user profile. SONAR [147] uses cosine similarity as a clustering task for attack detection. More advanced techniques are used in other works, which generally fall under the umbrella of ML. Traditional ML, neural networks (NN), and DL are all subsets of artificial intelligence (AI), as seen in Figure 10. They aim to train machines to behave like humans but differ in their learning models. For example, DataFreq [150] uses Logistic Regression (LR) for supervised ML to classify tweet sentiment as positive or negative. DeepNN [118] uses DL with CNN to classify tweets as relevant to security or not, with deeper learning as the number of layers increases.

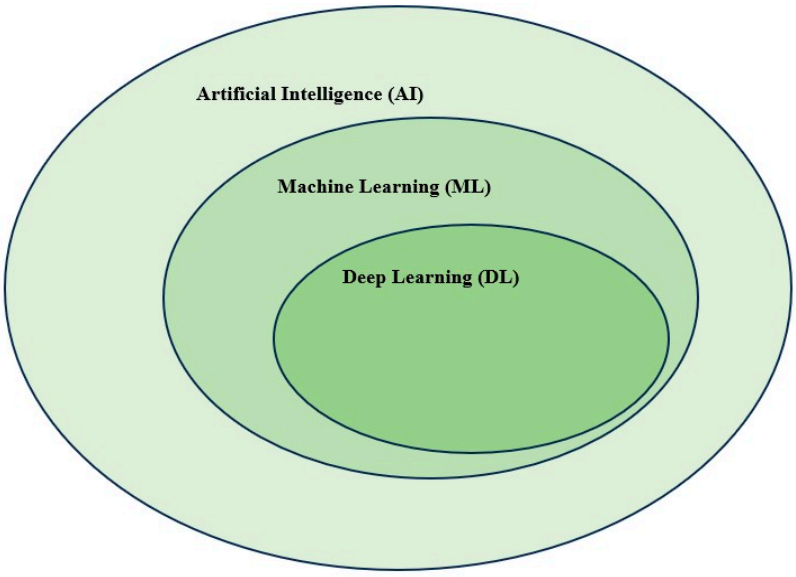

Figure 10. The complexity of algorithms AI, ML.

SYNAPSE utilizes Support Vector Machine (SVM) to predict tweet classification. While DL algorithms are highly sophisticated, they have the drawback of needing larger amounts of data to train the model and achieve desired outcomes. Table 4 organizes the studies according to the ascending complexity or accuracy of the algorithm used.

Table 4. Comparison of the intricacy of the algorithms.

Prediction Methods	Algorithm
DeepNN [118]	CNN
SYNAPSE [148]	SVM
DataFreq [150]	LR
SONAR [147]	Cosine similarity
CyberX [144]	SWRL
Text mining [146]	Filtering

7.2. Degree of Information Summarization

Due to the time-sensitive nature of cyber attacks, real-time detection is an essential feature. Our research aims to achieve this goal with varying degrees of information presented to the analyst. Providing a summary of detected events/attacks saves time and effort for the analyst, as the increased amount of information presented can prolong the analysis and necessary action. Summarization can be performed through various techniques, one of which is the clustering of security-related tweets. On one the one hand, the detection of an attack is made more dependable, and it stops people from starting an attack event. Nonetheless, this volume of data may not be fitting for security areas like SONAR [147]. Moreover, alerts may be used to restrict data. For instance, CyberX [144] sends an alert and presents all relevant tweets based on the user's profile. Named Entity Recognition (NER) is used in DeepNN [118] for each tweet by means of NN, which enhances the quality of data extraction, even though the number of tweets is still high. In DataFreq, the average sentiment of each company is displayed through a user-friendly interface. Clustering may be trailed by exemplar extraction, leading to a tweet representing each cluster, such as SYNAPSE's work [148]. Text mining [146] showcases the discovered attack term, frequency, and context as the final outcome. Table 5 outlines the comparison of the studies in terms of the amount of information presented to the user. The typical process involves clustering comparable tweets, extracting an exemplar for each cluster, and implementing the NER phase to extract entities from the tweet. By doing this, security analysts will have the most critical and useful data to comprehend the security hazard.

Table 5. Comparison of degree of information summarization.

Prediction Methods	Summarization
Text mining [146]	Summarized alert
SYNAPSE [148]	Clustering, exemplar
DataFreq [150]	Sentiment score for each company
DeepNN [118]	Classification, NER
CyberX [144]	Detailed alert
SONAR [147]	Clustering

7.3. Scalability and Effectiveness

This section examines and compares security research schemes that are based on essential time properties, as demonstrated in Table 6. The research focuses on the rapidly evolving nature of security terminologies and specific attack types. To avoid the problem of manually collecting keywords, which can lead to forgetting certain words, researchers suggest searching for new words to update the list of security keywords used in X searches. For example, the SONAR [147] system automatically discovers new related words and

allows users to evaluate them manually. If the new words are relevant, they can be added to the list for future use. The GloVe word embedding technique is used to extract semantic relationships between words, which helps the model keep up with changes in the field. However, the system still relies on human decision making, which can lead to mistakes. The researchers caution that the automatic addition of newly discovered words could increase the model's false positive rate. The study DataFreq [150] utilized TF and TF-IDF analysis to identify relevant keywords in retrieved tweets. While both methods focused on updating the keyword list, TF-IDF relied solely on word frequency and did not extract semantic features. The evaluation of X accounts used to retrieve the tweets was not fully assessed, hence a percentage was not assigned. Other studies [144,146], did not consider this aspect and followed previous research. It is assumed that users manually added new keywords to the list and received a 50% rating for doing so. This factor examines the metrics used to assess each study. In the security field, detecting all attacks is crucial, and detecting false positives is more acceptable than missing important alerts. The security attack detection model is a high recall model, as it measures the true positive rate (TPR) or the number of attacks detected by the model compared to the total actual attacks. Studies using ML algorithms can be compared using the same metric, with and achieving a TPR higher than 90, considered a reasonable degree of recall. DataFreq achieved a recall of about 84% and a precision of 85%, consistent with previous works. Studies not using ML such as, [144], and evaluated accuracy by the quality of generated alerts, similar to precision, where the number of true detected attacks is divided by the total alerts generated. Cyber-X Mittal, Das, Mulwad, Joshi and Finin [144] had the highest precision at 86%, even with the label "maybe" considered negative. Text mining [146] calculated an average evaluation by five annotators, with 84% correct detected terms. SONAR [147] had only 23 relevant detected events out of 100 in the evaluation period. The studies are arranged from best to worst performance in Table 7.

Table 6. Comparison of scalability.

Prediction Methods	Scalability
SONAR [147]	Updated keywords
DataFreq [150]	Updated keywords
SYNAPSE [148]	Fixed keywords and accounts
DeepNN [118]	Fixed keywords and accounts
Text mining [146]	Fixed accounts and dictionaries
CyberX [144]	Fixed user profile

Table 7. Evaluation of the effectiveness of models.

Prediction Methods	Recall (TPR)/Precision
DeepNN [118]	Recall 94%
SYNAPSE [148]	Recall 90%
DataFreq [150]	Recall 84%
CyberX [144]	Precision 86%
Text mining [146]	Precision 84%
SONAR [147]	Precision 23%

7.4. Semantic Characteristics

Within this aspect, Table 8 compares different studies based on the type of feature used for classification, specifically whether semantic features or keyword, count, or frequency features were utilized. The focus was on the accuracy of the technique, with emphasis on the positive effects of including semantic features as they extract both the content and meaning of the tweet. Some studies, such as text mining [146], did not use semantic features, instead utilizing dictionaries to filter unfamiliar terms. Others, such as [144,147], used keyword-based X searches to identify security-related tweets. DataFreq used n-gram to extract sentiment while SYNAPSE used TF-IDF to transform tweets into numerical values. However, simple textual and frequency-based feature extraction solutions may not accurately represent subtle semantic differences between real and false events mentioned. In contrast, the word2vec technique used by DeepNN transforms each tweet into vectors and can effectively detect similarities between words. Based on this, studies were arranged from the most semantic based to the least and given a percentage score. Even the best technique was given a 90% score as it could be replaced by char-based feature extraction techniques.

Table 8. Comparison of semantic characteristics.

Prediction Methods	Utilized Characteristics
DeepNN [118]	Word2vec
SYNAPSE [148]	TF-IDF
DataFreq [150]	N-gram
SONAR [147]	Keyword based
CyberX [144]	Keyword based
Text mining [146]	Simple filtering

7.5. The Scope of Detecting a Threat

A security discussion is any tweet containing a security-related keyword, while a security attack tweet must mention an imminent or ongoing attack. Finally, an IT infrastructure security attack tweet must contain a security keyword and mention an attack on a specific IT infrastructure. By limiting our scope to IT infrastructure attacks, we expect to improve the accuracy of our detection. This approach has been supported by previous studies such as SONAR, which found that detection based solely on keyword presence led to a 25% relevant detection rate. To further refine our results, we will exclude general discussions and tweets that mention security keywords but do not relate to an attack. Other studies [144], and have taken a similar approach, but focused on specific IT infrastructures. Our detection rate will begin at 50% for security events, increase to 75% for any security attack, and reach 100% for attacks on specific IT infrastructures, as shown in Table 9.

Table 9. Comparison of the scope of detection.

Prediction Methods	Detection Range
SYNAPSE [148]	IT security attack
DeepNN [118]	IT security attack
DataFreq [150]	IT security attack
CyberX [144]	IT security attack
Text mining [146]	Security attack
SONAR [147]	Security events

8. Discussion and Potential Opportunities

This discussion offers an in-depth comparison of six studies previously mentioned in the section focused on security content detection: SYNAPSE, DeepNN, DataFreq, CyberX [144], text mining, and SONAR. The SONAR model leverages cosine similarity for its predictions. Despite the common use of cosine similarity in text and vector space analysis, the model exhibits a relatively low precision of 23%. This suggests that while the model can identify true positives, it also generates a significant number of false positives. Consequently, SONAR might be better suited for applications where the primary goal is recall rather than precision, or where further filtering steps can be applied to reduce false positives. The text-mining [146] approach uses filtering techniques to achieve a precision of 84%. This high precision indicates that the model effectively minimizes false positives. Filtering techniques often involve rules or patterns to exclude irrelevant or erroneous data, making this model suitable for applications requiring high accuracy in positive predictions, such as specific keyword extraction or targeted information retrieval. The CyberX [144] model utilizes SWRL to achieve a precision of 86%. SWRL is a powerful tool for representing and reasoning with rules on the Semantic Web. The high precision rate indicates the model's effectiveness in applying semantic rules to filter and identify relevant information accurately. This makes CyberX particularly useful for scenarios requiring precise data extraction from large datasets, such as monitoring cyber threats or sentiment analysis. DataFreq [150] employs Logistic Regression to achieve a recall of 84%. Logistic regression is a statistical method commonly used for binary classification problems. A recall of 84% suggests that the model effectively identifies a large proportion of true positives. This makes it useful in applications where missing true positives is costly, such as in fraud detection or medical diagnosis. SYNAPSE [148] utilizes an SVM to achieve a recall of 90%. SVMs are known for their robustness in high-dimensional spaces and their effectiveness in classification tasks. A recall of 90% indicates the model's proficiency in capturing true positives, making it ideal for critical applications where it is essential to identify as many relevant instances as possible, such as in image recognition or bioinformatics.

In summary, the SONAR [147] model achieved a precision of 23% using the cosine similarity algorithm. This low precision suggests that SONAR's effectiveness in identifying relevant instances was relatively limited compared to other models. In contrast, text-mining [146] model demonstrated a significantly higher precision of 84% through filtering techniques. This indicates a more refined ability to accurately classify relevant data. CyberX [144] achieved an impressive precision of 86% utilizing SWRL. This performance highlights its robust capability in precise classification. DataFreq [150]: Focusing on recall, DataFreq attained a recall rate of 84% with Logistic Regression (LR). While its precision was not measured, the recall rate suggests a strong ability to identify relevant instances. SYNAPSE Alves et al. (2021) achieved a higher recall of 90% using Support Vector Machines (SVM). This indicates a superior capacity to retrieve relevant instances compared to DataFreq. DeepNN model reached an even higher recall of 94% through convolutional neural networks (CNN). This suggests an excellent performance in identifying relevant data points, Figure 11 presents the model's performance. By examining the total and average values of all factors combined and considering how the requirements and factors balance out based on their sum or average, we can see that DeepNN outperforms other techniques in both total and average figure of merit. While there are multiple techniques for char-based feature extraction available, we suggest taking a more semantic-based approach for this step. Additionally, the authors utilized a traditional SVM algorithm for classification. We propose using a DL algorithm with a deeper understanding of the data as an improvement, which we believe will enhance performance.

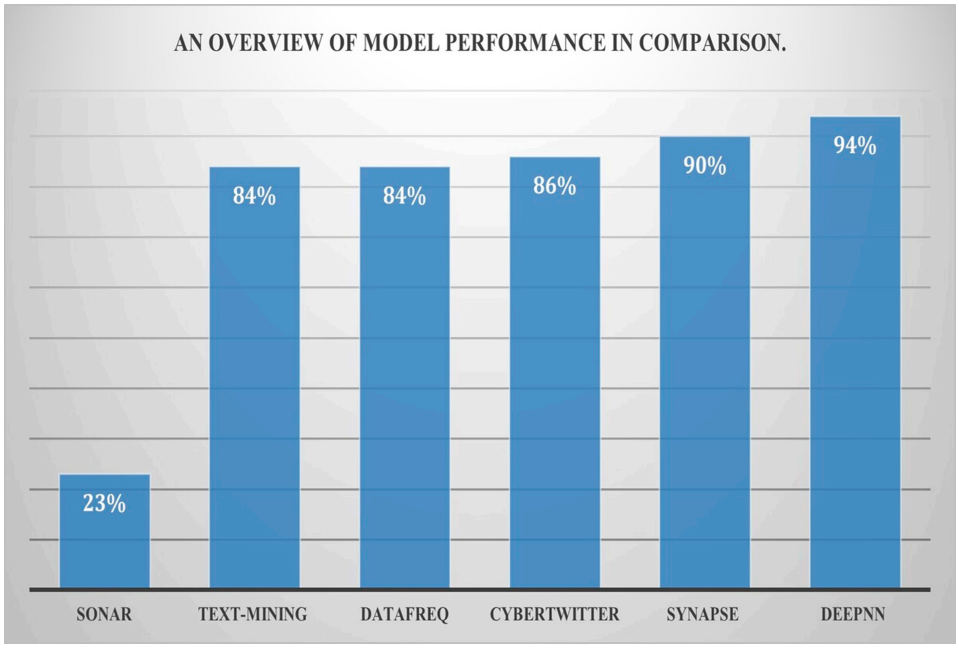

Figure 11. An overview of performance comparison.

9. Results and Analysis

According to the findings of this research study, there are several potential opportunities for enhancing and improving the current state of threat detection and prediction systems. One of the opportunities is to explore DL algorithms, which can potentially provide better performance than traditional ML algorithms like SVM and LR. Another opportunity is to adopt a semantic-based approach for feature extraction, which may enhance the accuracy and performance of the system. The integration of sentiment extraction for security breach identification is also recommended. The study suggests substituting LR with NN for sentiment extraction, which can potentially improve the accuracy and effectiveness of threat detection systems. Additionally, organizations can explore implementing DeepNN by [118] for system implementation and enhancement, as it has high ratings across most criteria. Further improvements can be made by substituting LR with NN for sentiment extraction and incorporating security breach identification. The study also recommends exploring different techniques and algorithms for feature extraction, such as a semantic-based approach, and using DL algorithms for classification. By combining the best performing techniques across all factors, there is an opportunity to develop a comprehensive threat detection system that can effectively detect and alert potential security breaches.

Overall, these potential opportunities can lead to the development of more accurate, efficient, and effective threat detection and prediction systems, which can help organizations to better protect their assets and systems from cyber threats.

10. Conclusions

Due to the increasing number of security attacks in recent years, monitoring and detecting attacks have become crucial for any organization. Different studies have applied various techniques to detect cybersecurity threats in X, including ML, natural language processing, and social network analysis. They have also varied in their data collection, classification approach, and evaluation metrics, which affects the accuracy of their analysis. More research is needed to develop accurate and effective detection techniques, and appropriate evaluation metrics are crucial to measure their performance. This study aimed to analyze and compare various attempts at utilizing X streaming data in real time to extract

knowledge about current and upcoming security attacks. The study examined different factors related to cybersecurity strategies that affect real-life situations, including scope of detecting, the investigation of semantic characteristics, the complexity of the algorithm, degree of information condensation, scalability, and effectiveness. This study proposed multiple improvements to maximize the benefits of security information published on X. Our study showed that the DeepNN by [118] consistently outperforms competing methods in terms of overall and average figure of merit. While character-level feature extraction methods are abundant, we contend that a semantic focus is more beneficial for this stage of the process. These feedback improvements are crucial as no study has achieved the best results in all factors, indicating a need for further research. We suggest future research directions and using deeper learning work in this area. Ultimately, it is all about providing an overview of cybersecurity challenges, deep learning solutions, and potential opportunities on the X platform. Based on our Survey, we recommend further research and solutions for cyber-threat detection on X using deep learning techniques.

Author Contributions: Conceptualization, O.A. and X.Z.; methodology, O.A. and X.Z.; software, O.A.; validation, O.A. and X.Z.; formal analysis, O.A., X.Z., R.G., A.S. and E.B.; investigation, O.A. and X.Z.; resources, O.A.; data curation, O.A.; writing—original draft preparation, O.A.; writing—review and editing, O.A., X.Z. and E.B.; supervision X.Z., R.G. and A.S. All authors have read and agreed to the published version of this manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We extend our gratitude to the anonymous reviewers whose insightful comments and suggestions significantly enhanced and clarified this manuscript.

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

References

1. Familoni, B.T. Cybersecurity challenges in the age of AI: Theoretical approaches and practical solutions. *Comput. Sci. IT Res. J.* **2024**, *5*, 703–724. [CrossRef]
2. Statista. Worldwide Cybersecurity Spending 2017–2028, Statista. 2024. Available online: https://www.statista.com/statistics/991304/worldwide-cybersecurity-spending/ (accessed on 10 November 2023).
3. Aiyer, B.; Caso, J.; Russell, P.; Sorel, M. New survey reveals $2 trillion market opportunity for cybersecurity technology and service providers. *Governance* **2022**, *1*, 2.
4. Kaur, G.; Bonde, U.; Pise, K.L.; Yewale, S.; Agrawal, P.; Shobhane, P.; Maheshwari, S.; Pinjarkar, L.; Gangarde, R. Social Media in the Digital Age: A Comprehensive Review of Impacts, Challenges and Cybercrime. *Eng. Proc.* **2024**, *62*, 6. [CrossRef]
5. Boyd, D.M.; Ellison, N.B. Social network sites: Definition, history, and scholarship. *J. Comput.-Mediat. Commun.* **2007**, *13*, 210–230.
6. Weir, G.R.; Toolan, F.; Smeed, D. The threats of social networking: Old wine in new bottles? *Inf. Secur. Tech. Rep.* **2011**, *16*, 38–43. [CrossRef]
7. Zigomitros, A.; Papageorgiou, A.; Patsakis, C. Social network content management through watermarking. In Proceedings of the 2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications, Liverpool, UK, 25–27 June 2012; pp. 1381–1386.
8. Stokes, K.; Carlsson, N. A peer-to-peer agent community for digital oblivion in online social networks. In Proceedings of the 2013 Eleventh Annual Conference on Privacy, Security and Trust, Tarragona, Spain, 10–12 July 2013; pp. 103–110.
9. Miller, Z.; Dickinson, B.; Deitrick, W.; Hu, W.; Wang, A.H. Twitter spammer detection using data stream clustering. *Inf. Sci.* **2014**, *260*, 64–73. [CrossRef]
10. Joe, M.M.; Ramakrishnan, B. Novel authentication procedures for preventing unauthorized access in social networks. *Peer-to-Peer Netw. Appl.* **2017**, *10*, 833–843.

11. Ghazinour, K.; Matwin, S.; Sokolova, M. YOURPRIVACYPROTECTOR, A recommender system for privacy settings in social networks. *arXiv* **2016**, arXiv:1602.01937.
12. Tounsi, W.; Rais, H. A survey on technical threat intelligence in the age of sophisticated cyber attacks. *Comput. Secur.* **2018**, *72*, 212–233. [CrossRef]
13. De Souza, G.A.; Da Costa-Abreu, M. Automatic offensive language detection from Twitter data using machine learning and feature selection of metadata. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–6.
14. Fang, Y.; Gao, J.; Liu, Z.; Huang, C. Detecting Cyber Threat Event from Twitter Using IDCNN and BiLSTM. *Appl. Sci.* **2020**, *10*, 5922. [CrossRef]
15. Humayun, M.; Niazi, M.; Jhanjhi, N.; Alshayeb, M.; Mahmood, S. Cyber Security Threats and Vulnerabilities: A Systematic Mapping Study. *Arab. J. Sci. Eng.* **2020**, *45*, 3171–3189. [CrossRef]
16. Dionísio, N.; Alves, F.; Ferreira, P.M.; Bessani, A. Towards end-to-end cyberthreat detection from Twitter using multi-task learning. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8.
17. Oosthoek, K.; Doerr, C. Cyber Threat Intelligence: A Product Without a Process? *Int. J. Intell. CounterIntell.* **2020**, *34*, 300–315. [CrossRef]
18. Rathore, S.; Sharma, P.K.; Loia, V.; Jeong, Y.-S.; Park, J.H. Social network security: Issues, challenges, threats, and solutions. *Inf. Sci.* **2017**, *421*, 43–69.
19. de Andrade, N.N.G.; Martin, A.; Monteleone, S. "All the better to see you with, my dear": Facial recognition and privacy in online social networks. *IEEE Secur. Priv.* **2013**, *11*, 21–28. [CrossRef]
20. González-Manzano, L.; González-Tablas, A.I.; de Fuentes, J.M.; Ribagorda, A. Cooped: Co-owned personal data management. *Comput. Secur.* **2014**, *47*, 41–65. [CrossRef]
21. Viejo, A.; Castella-Roca, J.; Rufián, G. Preserving the user's privacy in social networking sites. In Proceedings of the International Conference on Trust, Privacy and Security in Digital Business, Prague, Czech Republic, 28–29 August 2013; pp. 62–73.
22. Van Laere, O.; Schockaert, S.; Dhoedt, B. Georeferencing Flickr resources based on textual meta-data. *Inf. Sci.* **2013**, *238*, 52–74.
23. Lee, S.; Kim, J. Warningbird: A near real-time detection system for suspicious urls in twitter stream. *IEEE Trans. Dependable Secur. Comput.* **2013**, *10*, 183–195. [CrossRef]
24. Ahmed, F.; Abulaish, M. A generic statistical approach for spam detection in online social networks. *Comput. Commun.* **2013**, *36*, 1120–1129. [CrossRef]
25. Singh, S.; Jeong, Y.-S.; Park, J.H. A survey on cloud computing security: Issues, threats, and solutions. *J. Netw. Comput. Appl.* **2016**, *75*, 200–222. [CrossRef]
26. Squicciarini, A.C.; Shehab, M.; Wede, J. Privacy policies for shared content in social network sites. *VLDB J.* **2010**, *19*, 777–796.
27. Ramzan, N.; Park, H.; Izquierdo, E. Video streaming over P2P networks: Challenges and opportunities. *Signal Process. Image Commun.* **2012**, *27*, 401–411. [CrossRef]
28. Gurunath, R.; Klaib, M.F.J.; Samanta, D.; Khan, M.Z. Social media and steganography: Use, risks and current status. *IEEE Access* **2021**, *9*, 153656–153665.
29. Alsodi, O.; Zhou, X.; Gururajan, R.; Shrestha, A. A Survey on Detection of cybersecurity threats on Twitter using deep learning. In Proceedings of the 2021 8th International Conference on Behavioral and Social Computing (BESC), Doha, Qatar, 29–31 October 2021; pp. 1–5.
30. Zhang, Z.; Gupta, B.B. Social media security and trustworthiness: Overview and new direction. *Future Gener. Comput. Syst.* **2018**, *86*, 914–925.
31. Nauman, M.; Azam, N.; Yao, J. A three-way decision making approach to malware analysis using probabilistic rough sets. *Inf. Sci.* **2016**, *374*, 193–209.
32. Faghani, M.R.; Saidi, H. Malware propagation in online social networks. In Proceedings of the 2009 4th International Conference on Malicious and Unwanted Software (MALWARE), Montreal, QC, Canada, 13–14 October 2009; pp. 8–14.
33. Lanza, C.; Lodi, L. Towards a semi-automatic classifier of malware through tweets for early warning threat detection. *JLIS. It* **2024**, *15*, 101–118.
34. Noh, G.; Oh, H.; Kang, Y.-M.; Kim, C.-K. PSD: Practical Sybil detection schemes using stickiness and persistence in online recommender systems. *Inf. Sci.* **2014**, *281*, 66–84.
35. Faghani, M.R.; Nguyen, U.T. A study of clickjacking worm propagation in online social networks. In Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014), Redwood City, CA, USA, 13–15 August 2014; pp. 68–73.
36. Krombholz, K.; Hobel, H.; Huber, M.; Weippl, E. Advanced social engineering attacks. *J. Inf. Secur. Appl.* **2015**, *22*, 113–122.
37. Shah, A.; Varshney, S.; Mehrotra, M. Threats on online social network platforms: Classification, detection, and prevention techniques. *Multimed. Tools Appl.* **2024**, 1–33. [CrossRef]

38. Diomidous, M.; Chardalias, K.; Magita, A.; Koutonias, P.; Panagiotopoulou, P.; Mantas, J. Social and psychological effects of the internet use. *Acta Inform. Med.* **2016**, *24*, 66. [CrossRef]
39. El Asam, A.; Samara, M. Cyberbullying and the law: A review of psychological and legal challenges. *Comput. Hum. Behav.* **2016**, *65*, 127–141. [CrossRef]
40. Dreßing, H.; Bailer, J.; Anders, A.; Wagner, H.; Gallas, C. Cyberstalking in a large sample of social network users: Prevalence, characteristics, and impact upon victims. *Cyberpsychol. Behav. Soc. Netw.* **2014**, *17*, 61–67.
41. Munk, T. *The Rise of Politically Motivated Cyber Attacks: Actors, Attacks and Cybersecurity*; Routledge: London, UK, 2022.
42. Akoto, W. Who spies on whom? Unravelling the puzzle of state-sponsored cyber economic espionage. *J. Peace Res.* **2024**, *61*, 59–71.
43. Graham, C.M.; Lu, Y. Skills expectations in cybersecurity: Semantic network analysis of job advertisements. *J. Comput. Inf. Syst.* **2023**, *63*, 937–949. [CrossRef]
44. Dawson, M.E., Jr. *Cyber Warfare: Threats and Opportunities*; Postdoctoral report; Universidade Fernando Pessoa: Porto, Portugal, 2021.
45. Gadekar, C.; Rakshit, P.P. Study to Perform Opinion Mining on Motivation Factors Generating Cyber Crime by Twitter Analytics. In Proceedings of the International Conference on Innovative Computing & Communications (ICICC), New Delhi, India, 21–23 February 2020.
46. Romagna, M.; Leukfeldt, R.E. Social Opportunity Structures in Hacktivism: Exploring Online and Offline Social Ties and the Role of Offender Convergence Settings in Hacktivist Networks. *Vict. Offenders* **2024**, *19*, 511–533. [CrossRef]
47. Manakitsa, N.; Maraslidis, G.S.; Moysis, L.; Fragulis, G.F. A review of machine learning and deep learning for object detection, semantic segmentation, and human action recognition in machine and robotic vision. *Technologies* **2024**, *12*, 15. [CrossRef]
48. Lughbi, H.; Mars, M.; Almotairi, K. A Novel NLP-Driven Dashboard for Interactive CyberAttacks Tweet Classification and Visualization. *Information* **2024**, *15*, 137. [CrossRef]
49. Btoush, E.A.L.M.; Zhou, X.; Gururajan, R.; Chan, K.C.; Genrich, R.; Sankaran, P. A systematic review of literature on credit card cyber fraud detection using machine and deep learning. *PeerJ Comput. Sci.* **2023**, *9*, e1278.
50. Omar, S.; Ngadi, A.; Jebur, H.H. Machine learning techniques for anomaly detection: An overview. *Int. J. Comput. Appl.* **2013**, *79*, 33–41.
51. Zhang, C.; Lu, Y. Study on artificial intelligence: The state of the art and future prospects. *J. Ind. Inf. Integr.* **2021**, *23*, 100224.
52. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444.
53. Ayodele, T.O. Machine learning overview. *New Adv. Mach. Learn.* **2010**, *2*, 16.
54. Kotsiantis, S.B.; Zaharakis, I.; Pintelas, P. Supervised machine learning: A review of classification techniques. *Emerg. Artif. Intell. Appl. Comput. Eng.* **2007**, *160*, 3–24.
55. Hatcher, W.G.; Yu, W. A survey of deep learning: Platforms, applications and emerging research trends. *IEEE Access* **2018**, *6*, 24411–24432.
56. Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; Flammini, A. The rise of social bots. *Commun. ACM* **2016**, *59*, 96–104.
57. Varol, O.; Ferrara, E.; Davis, C.; Menczer, F.; Flammini, A. Online human-bot interactions: Detection, estimation, and characterization. In Proceedings of the Eleventh International AAAI Conference on Web and Social Media, Montreal, QC, Canada, 15–18 May 2017; pp. 280–289.
58. Lee, K.; Eoff, B.; Caverlee, J. Seven months with the devils: A long-term study of content polluters on twitter. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, 17–21 July 2011; pp. 185–192.
59. Kantepe, M.; Ganiz, M.C. Preprocessing framework for Twitter bot detection. In Proceedings of the 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, Turkey, 5–8 October 2017; pp. 630–634.
60. Subrahmanian, V.S.; Azaria, A.; Durst, S.; Kagan, V.; Galstyan, A.; Lerman, K.; Zhu, L.; Ferrara, E.; Flammini, A.; Menczer, F. The DARPA Twitter bot challenge. *Computer* **2016**, *49*, 38–46. [CrossRef]
61. David, I.; Siordia, O.S.; Moctezuma, D. Features combination for the detection of malicious Twitter accounts. In Proceedings of the 2016 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC), Ixtapa, Mexico, 9–11 November 2016; pp. 1–6.
62. Khaled, S.; El-Tazi, N.; Mokhtar, H.M. Detecting fake accounts on social media. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 3672–3681.
63. Yang, C.; Harkreader, R.; Gu, G. Empirical evaluation and new design for fighting evolving twitter spammers. *IEEE Trans. Inf. Forensics Secur.* **2013**, *8*, 1280–1293.
64. Velayutham, T.; Tiwari, P.K. Bot identification: Helping analysts for right data in twitter. In Proceedings of the 2017 3rd International Conference on Advances in Computing, Communication & Automation (ICACCA) (Fall), Dehradun, India, 15–16 September 2017; pp. 1–5.
65. Amleshwaram, A.A.; Reddy, N.; Yadav, S.; Gu, G.; Yang, C. Cats: Characterizing automation of twitter spammers. In Proceedings of the 2013 Fifth International Conference on Communication Systems and Networks (COMSNETS), Bangalore, India, 7–10 January 2013; pp. 1–10.

66. Ji, Y.; He, Y.; Jiang, X.; Cao, J.; Li, Q. Combating the evasion mechanisms of social bots. *Comput. Secur.* **2016**, *58*, 230–249.
67. Teljstedt, C.; Rosell, M.; Johansson, F. A semi-automatic approach for labeling large amounts of automated and non-automated social media user accounts. In Proceedings of the 2015 Second European Network Intelligence Conference, Karlskrona, Sweden, 21–22 September 2015; pp. 155–159.
68. Gilani, Z.; Kochmar, E.; Crowcroft, J. Classification of twitter accounts into automated agents and human users. In Proceedings of the ASONAM '17: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, Sydney, Australia, 31 July–3 August 2017; pp. 489–496.
69. Daouadi, K.E.; Rebaï, R.Z.; Amous, I. Bot detection on online social networks using deep forest. In Proceedings of the Artificial Intelligence Methods in Intelligent Algorithms: Proceedings of 8th Computer Science Online Conference 2019, Zlin, Czech Republic, 24–27 April 2019; pp. 307–315.
70. Yang, W.; Dong, G.; Wang, W.; Shen, G.; Gong, L.; Yu, M.; Lv, J.; Hu, Y. Detecting bots in follower markets. In Proceedings of the 9th International Conference, BIC-TA 2014, Wuhan, China, 16–19 October 2014; pp. 525–530.
71. Chu, Z.; Gianvecchio, S.; Wang, H.; Jajodia, S. Who is tweeting on Twitter: Human, bot, or cyborg? In Proceedings of the ACSAC '10: Proceedings of the 26th Annual Computer Security Applications Conference, Austin, TX, USA, 6–10 December 2010; pp. 21–30.
72. Chu, Z.; Gianvecchio, S.; Wang, H.; Jajodia, S. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Trans. Dependable Secur. Comput.* **2012**, *9*, 811–824.
73. Gurajala, S.; White, J.S.; Hudson, B.; Matthews, J.N. Fake Twitter accounts: Profile characteristics obtained using an activity-based pattern detection approach. In Proceedings of the SMSociety '15: Proceedings of the 2015 International Conference on Social Media & Society, Toronto, ON, Canada, 27–29 July 2015; pp. 1–7.
74. Caruccio, L.; Desiato, D.; Polese, G. Fake account identification in social networks. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 5078–5085.
75. Valliyammai, C.; Devakunchari, R. Distributed and scalable Sybil identification based on nearest neighbour approximation using big data analysis techniques. *Clust. Comput.* **2019**, *22* (Suppl. S6), 14461–14476.
76. Cai, C.; Li, L.; Zeng, D. Detecting social bots by jointly modeling deep behavior and content information. In Proceedings of the CIKM '17: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, 6–10 November 2017; pp. 1995–1998.
77. Kudugunta, S.; Ferrara, E. Deep neural networks for bot detection. *Inf. Sci.* **2018**, *467*, 312–322.
78. Wang, W.; Mauleon, R.; Hu, Z.; Chebotarov, D.; Tai, S.; Wu, Z.; Li, M.; Zheng, T.; Fuentes, R.R.; Zhang, F. Genomic variation in 3010 diverse accessions of Asian cultivated rice. *Nature* **2018**, *557*, 43–49. [CrossRef]
79. Ping, H.; Qin, S. A social bots detection model based on deep learning algorithm. In Proceedings of the 2018 IEEE 18th International Conference on Communication Technology (ICCT), Chongqing, China, 8–11 October 2018; pp. 1435–1439.
80. Morstatter, F.; Wu, L.; Nazer, T.H.; Carley, K.M.; Liu, H. A new approach to bot detection: Striking the balance between precision and recall. In Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, USA, 18–21 August 2016; pp. 533–540.
81. Igawa, R.A.; Barbon, S., Jr.; Paulo, K.C.S.; Kido, G.S.; Guido, R.C.; Júnior, M.L.P.; da Silva, I.N. Account classification in online social networks with LBCA and wavelets. *Inf. Sci.* **2016**, *332*, 72–83. [CrossRef]
82. Jr, S.B.; Campos, G.F.; Tavares, G.M.; Igawa, R.A.; Jr, M.L.P.; Guido, R.C. Detection of human, legitimate bot, and malicious bot in online social networks based on wavelets. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2018**, *14*, 1–17. [CrossRef]
83. Bara, I.-A.; Fung, C.J.; Dinh, T. Enhancing Twitter spam accounts discovery using cross-account pattern mining. In Proceedings of the 2015 IFIP/IEEE International Symposium on Integrated Network Management (IM), Ottawa, ON, Canada, 11–15 May 2015; pp. 491–496.
84. Gupta, A.; Budania, H.; Singh, P.; Singh, P.K. Facebook based choice filtering. In Proceedings of the 2017 IEEE 7th International Advance Computing Conference (IACC), Hyderabad, India, 5–7 January 2017; pp. 875–879.
85. Main, W.; Shekokhar, N. Twitterati identification system. *Procedia Comput. Sci.* **2015**, *45*, 32–41. [CrossRef]
86. Dickerson, J.P.; Kagan, V.; Subrahmanian, V. Using sentiment to detect bots on twitter: Are humans more opinionated than bots? In Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014), Beijing, China, 17–20 August 2014; pp. 620–627.
87. Loyola-González, O.; Monroy, R.; Rodríguez, J.; López-Cuevas, A.; Mata-Sánchez, J.I. Contrast pattern-based classification for bot detection on twitter. *IEEE Access* **2019**, *7*, 45800–45817. [CrossRef]
88. Andriotis, P.; Takasu, A. Emotional bots: Content-based spammer detection on social media. In Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 11–13 December 2018; pp. 1–8.
89. Beskow, D.M.; Carley, K.M. Its all in a name: Detecting and labeling bots by their name. *Comput. Math. Organ. Theory* **2019**, *25*, 24–35. [CrossRef]
90. Zhang, C.; Zhang, G.; Sun, S. A mixed unsupervised clustering-based intrusion detection model. In Proceedings of the 2009 Third International Conference on Genetic and Evolutionary Computing, Guilin, China, 14–17 October 2009; pp. 426–428.

91. Chavoshi, N.; Hamooni, H.; Mueen, A. Debot: Twitter Bot Detection via Warped Correlation. In *ICDM*; IEEE: Piscataway, NJ, USA, 2016; Volume 18, pp. 28–65.
92. Cresci, S.; Di Pietro, R.; Petrocchi, M.; Spognardi, A.; Tesconi, M. DNA-inspired online behavioral modeling and its application to spambot detection. *IEEE Intell. Syst.* **2016**, *31*, 58–64. [CrossRef]
93. Cresci, S.; Di Pietro, R.; Petrocchi, M.; Spognardi, A.; Tesconi, M. Social fingerprinting: Detection of spambot groups through DNA-inspired behavioral modeling. *IEEE Trans. Dependable Secur. Comput.* **2017**, *15*, 561–576. [CrossRef]
94. Minnich, A.; Chavoshi, N.; Koutra, D.; Mueen, A. BotWalk: Efficient adaptive exploration of Twitter bot networks. In Proceedings of the ASONAM '17: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, Sydney, Australia, 31 July–3 August 2017; pp. 467–474.
95. Chew, P.A. Searching for unknown unknowns: Unsupervised bot detection to defeat an adaptive adversary. In Proceedings of the 11th International Conference, SBP-BRiMS 2018, Washington, DC, USA, 10–13 July 2018; pp. 357–366.
96. Chen, Z.; Tanash, R.S.; Stoll, R.; Subramanian, D. Hunting malicious bots on twitter: An unsupervised approach. In Proceedings of the 9th International Conference, SocInfo 2017, Oxford, UK, 13–15 September 2017; pp. 501–510.
97. Munschauer, M.; Nguyen, C.T.; Sirokman, K.; Hartigan, C.R.; Hogstrom, L.; Engreitz, J.M.; Ulirsch, J.C.; Fulco, C.P.; Subramanian, V.; Chen, J. The NORAD lncRNA assembles a topoisomerase complex critical for genome stability. *Nature* **2018**, *561*, 132–136. [CrossRef]
98. Abu-El-Rub, N.; Mueen, A. Botcamp: Bot-driven interactions in social campaigns. In Proceedings of the WWW '19: The World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019; pp. 2529–2535.
99. Zhu, X.; Goldberg, A.B. Introduction to semi-supervised learning. *Synth. Lect. Artif. Intell. Mach. Learn.* **2009**, *3*, 1–130.
100. Chapelle, O.; Scholkopf, B.; Zien, A. Semi-supervised learning (Chapelle, O. et al., eds.; 2006) [book reviews]. *IEEE Trans. Neural Netw.* **2009**, *20*, 542. [CrossRef]
101. Shi, P.; Zhang, Z.; Choo, K.-K.R. Detecting malicious social bots based on clickstream sequences. *IEEE Access* **2019**, *7*, 28855–28862. [CrossRef]
102. Dorri, A.; Abadi, M.; Dadfarnia, M. Socialbothunter: Botnet detection in twitter-like social networking services using semi-supervised collective classification. In Proceedings of the 2018 IEEE 16th International Conference on Dependable, Autonomic and Secure Computing, 16th International Conference on Pervasive Intelligence and Computing, 4th International Conference on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech), Athens, Greece, 12–15 August 2018; pp. 496–503.
103. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
104. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
105. Xu, H.; Dong, M.; Zhu, D.; Kotov, A.; Carcone, A.I.; Naar-King, S. Text classification with topic-based word embedding and convolutional neural networks. In Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, Seattle, WA, USA, 2–5 October 2016; pp. 88–97.
106. Olabanjo, O.; Wusu, A.; Aigbokhan, E.; Olabanjo, O.; Afisi, O.; Akinnuwesi, B. A novel graph convolutional networks model for an intelligent network traffic analysis and classification. *Int. J. Inf. Technol.* **2024**, 1–13. [CrossRef]
107. 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.
108. Gundubogula, A.S. Enhancing Graph Convolutional Network with Label Propagation and Residual for Malware Detection. Master's Thesis, Wright State University, Dayton, OH, USA, 2023.
109. Khan, Z.; Khan, Z.; Lee, B.-G.; Kim, H.K.; Jeon, M. Graph neural networks based framework to analyze social media platforms for malicious user detection. *Appl. Soft Comput.* **2024**, *155*, 111416.
110. Simran, K.; Balakrishna, P.; Vinayakumar, R.; Soman, K. Deep Learning Approach for Enhanced Cyber Threat Indicators in Twitter Stream. In Proceedings of the 7th International Symposium, SSCC 2019, Trivandrum, India, 18–21 December 2019; pp. 135–145.
111. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681.
112. Elman, J.L. Finding structure in time. *Cogn. Sci.* **1990**, *14*, 179–211. [CrossRef]
113. Shin, H.-S.; Kwon, H.-Y.; Ryu, S.-J. A new text classification model based on contrastive word embedding for detecting cybersecurity intelligence in twitter. *Electronics* **2020**, *9*, 1527. [CrossRef]
114. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [PubMed]
115. Wang, J.-H.; Liu, T.-W.; Luo, X.; Wang, L. An LSTM approach to short text sentiment classification with word embeddings. In Proceedings of the 2018 Conference on Computational Linguistics and Speech Processing, Hsinchu, Taiwan, 4–5 October 2018; pp. 214–223.
116. Ding, Z.; Xia, R.; Yu, J.; Li, X.; Yang, J. Densely connected bidirectional lstm with applications to sentence classification. In Proceedings of the 7th CCF International Conference, NLPCC 2018, Hohhot, China, 26–30 August 2018; pp. 278–287.

117. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [PubMed]
118. Dionísio, N.; Alves, F.; Ferreira, P.M.; Bessani, A. Cyberthreat detection from twitter using deep neural networks. In Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19 July 2019; pp. 1–8.
119. Berahman, K.; Zhou, X.; Li, Y.; Gururajan, R.; Barua, P.; Acharya, R.; Chennakesavan, S.K. New Ensemble Deep Learning Model for Gynaecological Cancer Risk Prediction; Research Square, Australia. 2024. Available online: https://www.researchgate.net/publication/379506676_New_Ensemble_Deep_Learning_Model_for_Gynaecological_Cancer_Risk_Prediction (accessed on 10 November 2023).
120. Shukla, H.; Jagtap, N.; Patil, B. Enhanced Twitter bot detection using ensemble machine learning. In Proceedings of the 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 20–22 January 2021; pp. 930–936.
121. Shahnawaz Ahmad, M.; Mehraj Shah, S. Unsupervised ensemble based deep learning approach for attack detection in IoT network. *Concurr. Comput. Pract. Exp.* **2022**, *34*, e7338.
122. Khanday, A.M.U.D.; Rabani, S.T.; Khan, Q.R.; Malik, S.H. Detecting twitter hate speech in COVID-19 era using machine learning and ensemble learning techniques. *Int. J. Inf. Manag. Data Insights* **2022**, *2*, 100120.
123. Ahmad, R.; Alsmadi, I.; Alhamdani, W.; Tawalbeh, L.a. A deep learning ensemble approach to detecting unknown network attacks. *J. Inf. Secur. Appl.* **2022**, *67*, 103196.
124. Muneer, A.; Alwadain, A.; Ragab, M.G.; Alqushaibi, A. Cyberbullying detection on social media using stacking ensemble learning and enhanced BERT. *Information* **2023**, *14*, 467. [CrossRef]
125. Siddiqui, T.; Hina, S.; Asif, R.; Ahmed, S.; Ahmed, M. An ensemble approach for the identification and classification of crime tweets in the English language. *Comput. Sci. Inf. Technol.* **2023**, *4*, 149–159.
126. Arora, R.; Gupta, R.; Yadav, P. Utilizing Ensemble Learning to enhance the detection of Malicious URLs in the Twitter dataset. In Proceedings of the 2024 4th International Conference on Innovative Practices in Technology and Management (ICIPTM), Noida, India, 21–23 February 2024; pp. 1–6.
127. Krishna, T.V.S.; Krishna, T.S.R.; Kalime, S.; Krishna, C.V.M.; Neelima, S.; PBV, R.R. A novel ensemble approach for Twitter sentiment classification with ML and LSTM algorithms for real-time tweets analysis. *Indones. J. Electr. Eng. Comput. Sci.* **2024**, *34*, 1904–1914.
128. Alqahtani, A.F.; Ilyas, M. A Machine Learning Ensemble Model for the Detection of Cyberbullying. *arXiv* **2024**, arXiv:2402.12538.
129. Olaitan, O.L.; David, A.O.; Michael, O.A. Deep Learning Approach for Classification of Tweets in Detecting Cyber Truculent. *Adv. Res.* **2024**, *25*, 113–122.
130. Vaiyapuri, T.; Shankar, K.; Rajendran, S.; Kumar, S.; Gaur, V.; Gupta, D.; Alharbi, M. Automated cyberattack detection using optimal ensemble deep learning model. *Trans. Emerg. Telecommun. Technol.* **2024**, *35*, e4899.
131. Ruohonen, J.; Hyrynsalmi, S.; Leppänen, V. A mixed methods probe into the direct disclosure of software vulnerabilities. *Comput. Hum. Behav.* **2020**, *103*, 161–173.
132. Campiolo, R.; Santos, L.A.F.; Batista, D.M.; Gerosa, M.A. Evaluating the utilization of Twitter messages as a source of security alerts. In Proceedings of the 28th Annual ACM Symposium on Applied Computing, Coimbra, Portugal, 18–22 March 2013; pp. 942–943.
133. Sabottke, C.; Suciu, O.; Dumitraș, T. Vulnerability disclosure in the age of social media: Exploiting twitter for predicting real-world exploits. In Proceedings of the 24th {USENIX} Security Symposium ({USENIX} Security 15), Washington, DC, USA, 12–14 August 2015; pp. 1041–1056.
134. Trabelsi, S.; Plate, H.; Abida, A.; Aoun, M.M.B.; Zouaoui, A.; Missaoui, C.; Gharbi, S.; Ayari, A. Mining social networks for software vulnerabilities monitoring. In Proceedings of the 2015 7th International Conference on New Technologies, Mobility and Security (NTMS), Paris, France, 27–29 July 2015; pp. 1–7.
135. Kergl, D.; Roedler, R.; Rodosek, G.D. Detection of zero day exploits using real-time social media streams. In Proceedings of the Advances in Nature and Biologically Inspired Computing: Proceedings of the 7th World Congress on Nature and Biologically Inspired Computing (NaBIC2015), Pietermaritzburg, South Africa, 1–3 December 2015; pp. 405–416.
136. Queiroz, A.; Keegan, B.; Mtenzi, F. Predicting software vulnerability using security discussion in social media. In Proceedings of the European Conference on Cyber Warfare and Security, Dublin, Ireland, 29–30 June 2017; pp. 628–634.
137. Behzadan, V.; Aguirre, C.; Bose, A.; Hsu, W. Corpus and deep learning classifier for collection of cyber threat indicators in twitter stream. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 5002–5007.
138. Arora, T.; Sharma, M.; Khatri, S.K. Detection of cyber crime on social media using random forest algorithm. In Proceedings of the 2019 2nd International Conference on Power Energy, Environment and Intelligent Control (PEEIC), Greater Noida, India, 18–19 October 2019; pp. 47–51.
139. Le, B.-D.; Wang, G.; Nasim, M.; Babar, M.A. Gathering cyber threat intelligence from Twitter using novelty classification. *arXiv* **2019**, arXiv:19f07.01755.

140. Mahaini, M.I.; Li, S. Detecting cyber security related Twitter accounts and different sub-groups: A multi-classifier approach. In Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Virtual, 8–11 November 2021; pp. 599–606.
141. Deshmukh, R.; Shinde, S.; Yadav, B.; Pathak, A.; Shetty, A. Darkintellect: An Approach to Detect Cyber Threat Using Machine Learning Techniques on Open-Source Information. *Math. Stat. Eng. Appl.* **2022**, *71*, 1431–1439.
142. Coyac-Torres, J.E.; Sidorov, G.; Aguirre-Anaya, E.; Hernández-Oregón, G. Cyberattack detection in social network messages based on convolutional neural networks and NLP techniques. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1132–1148. [CrossRef]
143. Abdelhaq, H.; Sengstock, C.; Gertz, M. Eventweet: Online localized event detection from twitter. *Proc. VLDB Endow.* **2013**, *6*, 1326–1329.
144. Mittal, S.; Das, P.K.; Mulwad, V.; Joshi, A.; Finin, T. Cybertwitter: Using twitter to generate alerts for cybersecurity threats and vulnerabilities. In Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, USA, 18–21 August 2016; pp. 860–867.
145. Horrocks, I.; Patel-Schneider, P.F.; Boley, H.; Tabet, S.; Grosof, B.; Dean, M. SWRL: A semantic web rule language combining OWL and RuleML. *W3C Memb. Submiss.* **2004**, *21*, 1–31.
146. Sapienza, A.; Bessi, A.; Damodaran, S.; Shakarian, P.; Lerman, K.; Ferrara, E. Early warnings of cyber threats in online discussions. In Proceedings of the 2017 IEEE International Conference on Data Mining Workshops (ICDMW), New Orleans, LA, USA, 18–21 November 2017; pp. 667–674.
147. Le Sceller, Q.; Karbab, E.B.; Debbabi, M.; Iqbal, F. Sonar: Automatic detection of cyber security events over the twitter stream. In Proceedings of the 12th International Conference on Availability, Reliability and Security, Reggio Calabria, Italy, 29 August–1 September 2017; pp. 1–11.
148. Alves, F.; Bettini, A.; Ferreira, P.M.; Bessani, A. Processing tweets for cybersecurity threat awareness. *Inf. Syst.* **2021**, *95*, 101586.
149. Nazir, F.; Ghazanfar, M.A.; Maqsood, M.; Aadil, F.; Rho, S.; Mehmood, I. Social media signal detection using tweets volume, hashtag, and sentiment analysis. *Multimed. Tools Appl.* **2019**, *78*, 3553–3586. [CrossRef]
150. Rodriguez, A.; Okamura, K. Generating real time cyber situational awareness information through social media data mining. In Proceedings of the 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), Milwaukee, WI, USA, 15–19 July 2019; pp. 502–507.
151. Dabiri, S.; Heaslip, K. Developing a Twitter-based traffic event detection model using deep learning architectures. *Expert Syst. Appl.* **2019**, *118*, 425–439.
152. Sani, A.M.; Moeini, A. Real-time Event Detection in Twitter: A Case Study. In Proceedings of the 2020 6th International Conference on Web Research (ICWR), Tehran, Iran, 22–23 April 2020; pp. 48–51.
153. Rodriguez, A.; Okamura, K. Enhancing data quality in real-time threat intelligence systems using machine learning. *Soc. Netw. Anal. Min.* **2020**, *10*, 1–22. [CrossRef]
154. Reddy, P.M.; Venkatesh, K.; Bhargav, D.; Sandhya, M. Spam detection and fake user identification methodologies in social networks using extreme machine learning. *Int. J. Anal. Exp. Modal Anal.* **2021**, *13*, 2367–2374. [CrossRef]
155. Kondeti, P.; Yerramreddy, L.P.; Pradhan, A.; Swain, G. Fake account detection using machine learning. In *Evolutionary Computing and Mobile Sustainable Networks: Proceedings of ICECMSN 2020, Proceedings of the International Conference on Evolutionary Computing and Mobile Sustainable Networks (ICECMSN 2020), Bangalore, India, 20–21 February 2020*; Springer: Singapore, 2021; pp. 791–802.
156. Bindu, K.; Rishith, B.P.; Sathish, D.; Subhash, V.; Harika, B.; Swathi, N. Detection of fake accounts in Twitter using data science. *Int. Res. J. Mod. Eng. Technol. Sci.* **2022**, *4*, 3552–3556.
157. Rodrigues, A.P.; Fernandes, R.; Shetty, A.; Lakshmanna, K.; Shafi, R.M. Real-time twitter spam detection and sentiment analysis using machine learning and deep learning techniques. *Comput. Intell. Neurosci.* **2022**, *2022*, 5211949. [CrossRef] [PubMed]
158. Shukla, R.; Sinha, A.; Chaudhary, A. TweezBot: An AI-driven online media bot identification algorithm for Twitter social networks. *Electronics* **2022**, *11*, 743. [CrossRef]
159. Mughaid, A.; Obeidat, I.; AlZu'bi, S.; Elsoud, E.A.; Alnajjar, A.; Alsoud, A.R.; Abualigah, L. A novel machine learning and face recognition technique for fake accounts detection system on cyber social networks. *Multimed. Tools Appl.* **2023**, *82*, 26353–26378. [CrossRef]
160. Ritter, A.; Wright, E.; Casey, W.; Mitchell, T. Weakly supervised extraction of computer security events from twitter. In Proceedings of the WWW '15: Proceedings of the 24th International Conference on World Wide Web, Florence, Italy, 18–22 May 2015; pp. 896–905.
161. Rao, P.; Kamhoua, C.; Njilla, L.; Kwiat, K. *Methods to Detect Cyberthreats on Twitter: 'Surveillance in Action'*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 333–350.
162. Chambers, N.; Fry, B.; McMasters, J. *Detecting Denial-of-Service Attacks from Social Media Text: Applying NLP to Computer Security*; Association for Computational Linguistics: New Orleans, LA, USA, 2018; pp. 1626–1635.
163. Yılmaz, Y.; Hero, A.O. Multimodal event detection in Twitter hashtag networks. *J. Signal Process. Syst.* **2018**, *90*, 185–200. [CrossRef]

164. Zong, S.; Ritter, A.; Mueller, G.; Wright, E. Analyzing the perceived severity of cybersecurity threats reported on social media. *arXiv* **2019**, arXiv:1902.10680.
165. Ghankutkar, S.; Sarkar, N.; Gajbhiye, P.; Yadav, S.; Kalbande, D.; Bakereywala, N. Modelling machine learning for analysing crime news. In Proceedings of the 2019 International Conference on Advances in Computing, Communication and Control (ICAC3), Mumbai, India, 20–21 December 2019; pp. 1–5.
166. Boyd, K.; Eng, K.H.; Page, C.D. Area under the precision-recall curve: Point estimates and confidence intervals. In Proceedings of the M16666achine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, 23–27 September 2013; Proceedings, Part III 13; pp. 451–466.