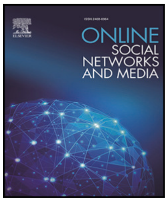
Online Social Networks and Media 30 (2022) 100211

Contents lists available at ScienceDirect

Online Social Networks and Media

journal homepage: www.elsevier.com/locate/osnem



A social network of crime: A review of the use of social networks for crime and the detection of crime

Brett Drury [c,d], Samuel Morais Drury [e], Md Arafatur Rahman [f], Ihsan Ullah [a,b,*]

[a] School of Computer Science, National University of Ireland Galway, Galway, Ireland
[b] CeADAR Ireland's Centre for Applied AI, University College Dublin, Dublin, Ireland
[c] LIAAD INESC Tec, INESC, Campus da FEUP, Porto, Portugal
[d] Liverpool Hope University, Hope Campus, Liverpool, UK
[e] Colégio Puríssimo R. Sete, 881 - Centro, Rio Claro, Brazil
[f] University Malaysia Pahang, Malaysia

ARTICLE INFO

Keywords:
Social media
Cybercrime
Machine learning
Crime prediction
NLP

ABSTRACT

Social media is used to commit and detect crimes. With automated methods, it is possible to scale both crime and detection of crime to a large number of people. The ability of criminals to reach large numbers of people has made this area subject to frequent study, and consequently, there have been several surveys that have reviewed specific crimes committed on social platforms. Until now, there has not been a review article that considers all types of crimes on social media, their similarity as well as their detection. The demonstration of similarity between crimes and their detection methods allows for the transfer of techniques and data between domains. This survey, therefore, seeks to document the crimes that have been committed on social media, and demonstrate their similarity through a taxonomy of crimes. Also, this survey documents publicly available datasets. Finally, this survey provides suggestions for further research in this field.

1. Introduction

Social media is becoming intertwined with peoples' day to day lives, where users post details of their lives which can be seen by their friends and the public. These posts may contain explicit and implicit details of a crime. This information can often be hidden from traditional crime statistics and the police in general because of the victim's reluctance to report a crime. This inertia may be due to: 1. Triviality of the crime, 2. The crime may be embarrassing to the individual and 3. The individual may not know that they have been a victim of a crime. The monitoring of social media may allow the relevant authorities to supplement traditional crime reporting. In common with many new technologies and methods of communication, social media is used to commit criminal acts. Social media gives criminals reach to individuals that was impossible before the invention and mass adoption of the Internet. Therefore, with little effort, criminals can commit crimes across legal jurisdictions against large numbers of people. Although social media can be used to commit a crime, it can also be used to detect and predict criminal acts. Criminals can leave traces of their crimes in posts, and in some cases, they openly boast about their actions. In addition, users can inadvertently leave predictors of crime in their posts, as well as reports of criminal actions that may be absent from traditional crime reports. Social media allows for the creation of novel crimes, however, these new crimes have a common root which could be described as a crime hyponym which gives these crimes a common modus operandi, although the aims of the crimes are different. The common manner in which different crimes are conducted infers that techniques from a similar crime could be successfully transferred to a newer crime that has a lack of data or existing detection techniques. Adaption of existing techniques to new crimes will allow cybersecurity researchers to react quickly to novel crimes on social media by quickly identifying similar crimes and well-established solutions. The motivation of this paper is to survey the broad use of social media to commit, detect and predict crime rather than concentrating on a single area of crime, which has been the main focus of survey papers in this area. The aim of this paper, therefore, is to 1. Demonstrate a link between information in social media and crime, 2. Centralise the broad body of research in this area, 3. Identify current research trends, 4. Document relevant resources, 5. Categorise social media into related groups, and 6. Suggest future directions of research.

[*] Corresponding author at: School of Computer Science, National University of Ireland Galway, Galway, Ireland.
E-mail address: ihsan.ullah@nuigalway.ie (I. Ullah).

https://doi.org/10.1016/j.osnem.2022.100211
Received 20 October 2021; Received in revised form 27 April 2022; Accepted 9 May 2022
Available online 27 May 2022

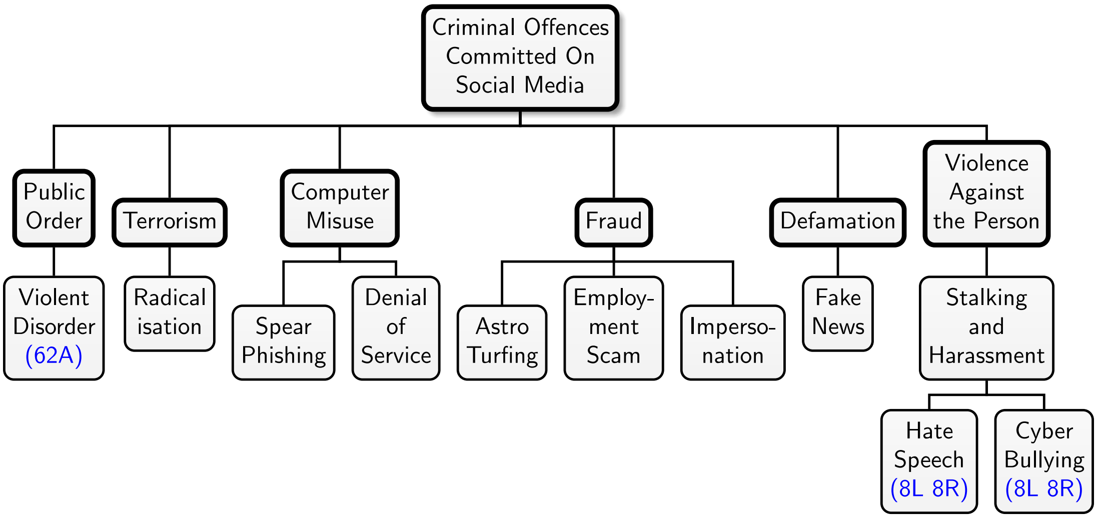


Fig. 1. Taxonomy of criminal offences committed on social media.

1.1. *Existing survey papers*

Crime and social media is a popular research area, consequently there are a number of survey papers that address this area. However, these survey papers tend to be targeted to a specific crime such as hate speech [1] and Spear Phishing [2], or the use of social media for policing in general [3]. To date, there has not been a survey that surveys the use of social media for crime and the connections between the crimes. The similarity of crimes and their detection allows for the repurposing of existing techniques to new areas and crimes.

1.2. *Structure of paper*

This paper is organised in the following order. Section 2 describes the classification of crime that is committed and detected on social media. In Section 3, the article discusses the crimes that are committed on social media, and their detection methods as well as frequently used learners and data representation methodologies. The techniques that use social media to extract information from users on social media to commit real-world crimes will be discussed in Section 4, and in Section 5 we will discuss the use of social media to detect real-world crimes and criminals. In Section 6, there will be details about tools and datasets that can be used to commit and detect crimes on social media. The survey will finally end with a conclusion and suggestions for future research in this area.

2. Classification of criminal offences committed on social media

Techniques that use social media to report criminal events or predict future crime are predicated upon one of the following assumptions:

- There is a direct causal link between information in social media and criminal acts
- There is a mechanism through which users disclose information that reports a crime or indicates a future crime, that they would not normally divulge to the relevant authorities.

The intuition that information in social media posts has a direct correlation to criminal acts is supported by [4–7]. Information published on social media is not limited to the prediction and reporting of physical crimes and can be applied to cybercrime and criminals [8].

In addition to being a proxy for crime statistics, social media can be used to commit crimes. Speech, and by extension social media posts, can be used to insult minority groups, manipulate and coerce individuals into committing crimes in the physical world.

This section will also describe a taxonomy of Criminal Offences. The subsections that follow the taxonomy will describe the general types of offences: public order, terrorism, computer misuse, fraud, defamation, and violence against the person.

2.1. *Morality of social media monitoring*

Mass surveillance of social media can produce some unease in users and critics. And there may be some criticism for projects that try to predict private information from publicly available information on social media and related applications [9]. Social media monitoring has been used for malevolent aims, for example, social media monitoring has also been used to identify vulnerable users and influence their voting behaviour [10]. These types of behaviour and actions by social media companies are generally seen as immoral, which is evidenced by the public and media reactions to the Cambridge Analytica and related scandals [11].

There is a moral case for social media monitoring, where the absence of monitoring and reporting may cause harm to users. This may be psychological harm or physical harm. The absence of monitoring for this type of user behaviour could be considered to be immoral. In addition, although currently social media companies are protected by safe harbour provisions [12], there has been, however, some momentum from lawmakers to make social media companies responsible for user activity and the content of their posts.[1] Therefore, it is argued that for the commission of crimes on social media, organisations have a moral imperative to monitor their platform and report criminal behaviour.

2.2. *Taxonomy of Criminal Offences*

Crimes committed on social media often use similar techniques, but are covered by different laws or a subset of a general offence. To aid the reader's understanding of the offences committed on social media, a simple taxonomy was developed, and it is shown in Fig. 1. The taxonomy is based upon UK Law and where possible the offences are aligned with the offence classification published by the UK government.

The motivation of this section is to discover crimes that are committed on social media, and where possible to group them under the relevant legislation. This will allow the identification of common motivations, which are documented in Fig. 1

The crimes found in the literature survey can often be variations of a common exploitation technique, and therefore approaches in one

[1] https://bit.ly/3h8aIDT.

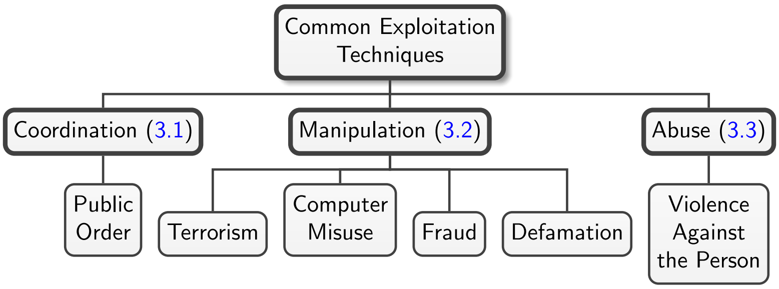


Fig. 2. Taxonomy of common social media exploitation techniques.

domain can be transferred to another with a reasonable expectation of success. To assist the reader, another taxonomy was developed to group together different offences under common exploitation techniques used on social media. This taxonomy is shown in Fig. 2. These taxonomies represent the connection of the crimes found in the literature review for this paper. A general classification of crimes and digital crimes found on social media is given by [13].

3. Criminal offences committed on social media

This section will discuss the use of social media to commit crimes, as well as the specific offences that are committed on social media. For this paper, three categories of the use of social media to commit crimes were defined. These are: coordination, manipulation and abuse, and the crimes that are committed with these techniques are grouped under them. This is because this paper hypothesises that strategies that detect crimes on social media can be repurposed to similar crimes under the same grouping.

The motivation of this section is to find common motivations for crimes that are committed on social media. It is a hypothesis of this paper that common motivations will have common methods to detect the associated crime

3.1. Coordination

Coordination is the use of social media to commit offences by organising one or more persons to commit a crime [14]. Social media may be used to incite or message individuals to commit crimes and have been used in recent events such as the insurrection[2] (United States Capitol attack 2021[3]) in the USA. It should be noted that using social media that simply incites violence on social media is sufficient for a custodial sentence in the UK.[4] The incitement does not have to result in violence or public disorder.

3.1.1. Public Order offences

Public Order offences can cover a multitude of offences such as violent disorder, unlawful assembly, affray and violent disorder. Public Order offences in the UK are covered by the Public Disorder Act of 1986. The offence that is committed on social media when planning public disorder acts, depending on the legal jurisdiction, is the offence of conspiracy and is referred to in the UK's legal system as an inchoate offence [15]. Inchoate offences not only cover conspiracy to commit public order offences, but individuals and groups who aid and assist public disorder acts. The literature review found that violent disorder was the only crime that could be categorised under public order offences.

3.1.1.1. Violent disorder. The definition of violent disorder in English law is given by the Public Order Act of 1986. It states a violent disorder offence is committed when "Where three or more persons who are present together use or threaten unlawful violence and the conduct of them (taken together) is such as would cause a person of reasonable firmness present at the scene to fear for his personal safety".[5] This rather broad definition will capture public disorder that does not have criminal intentions, such as political protests e.g. Arab Spring.

Violent disorder is not caused by social media, but it can propagate through or be organised on social media. Also, the advent of social media on mobile devices has encouraged the use of social media to enable violent disorder. This was the case in the Arab Spring protests[6] in 2011 [16].

The cause of the violent disorder was economic shocks to the populations of several Arab countries [17]. The use of social media to organise violent disorder is not limited to the Arab Spring, and has been used in protests in Iran, China [18], England [19], and Canada [20]. The use of social media to detect violent disorder is discussed in Section 5. But this small selection of papers demonstrates that it is possible to commit conspiracy to commit violent disorder on social media.

3.2. Manipulation

Manipulation exploits are typically when an attacker (criminal) seeks to manipulate a target (victim) into doing something they would not normally do. How an attacker manipulates the target is through language. Language, either as posts or direct messages, seeks to manipulate the actions of the target in the real or virtual world. There are five offences attributed to the manipulation exploit type: 1. Spear Phishing, 2. Employment Scams, 3. Radicalisation, 4. Astroturfing and 5. Fake News. Spear-phishing attacks are specific attacks against individuals into revealing secret information, such as bank account information. Employment Scams are frauds that use fake employment offers to manipulate the target into committing crimes. Radicalisation is a manipulation process that encourages the target to commit violent acts. Astroturfing and fake news are techniques that provide a false narrative about a subject or individual.

3.2.1. Computer Misuse

In the United Kingdom, the Computer Misuse Act of 1990 defines a series of offences that are related to personal data held by public and private organisations [21]. It should be noted that although this is UK legislation, it has an extraterritorial effect [21], and therefore the legislation has worldwide jurisdiction.

There is one crime, Spear Phishing, that is grouped under this general offence. Spear Phishing is an offence under the Computer Misuse act, as it seeks to gain unauthorised access to a computer system.

[2] https://archive.fo/msKNn.
[3] https://en.wikipedia.org/wiki/2021_United_States_Capitol_attack.
[4] https://www.bbc.co.uk/news/uk-england-manchester-14551582.
[5] https://archive.fo/qQOIt.
[6] It is the opinion of the authors that the Arab Spring qualifies as violent disorder under UK legislation. It should be noted that the UK legislation is generally regarded as very restrictive, and it has been used to suppress political dissent in the UK, such as the Poll Tax Demonstrations.

3.2.1.1. Spear Phishing. Spear Phishing attacks on specific users with a presence on social media can be automated using machine learning techniques such as Neural Networks [22]. Automated Spear Phishing has several distinct phases which can be represented by '5C' that represents: "Collect, Construct, Contact, Compromise, and Contagion" [23]. The collect phase is where the attacker gathers information upon the intended targets [23]. The data gathering can be achieved by using the associated API and keywords to identify suitable profiles and their tweets that leak personal information [23]. The contact phase is typically from an account that has followed or friended the target account. The contact messages are often automatically generated from the public content generated by the target account [23]. The message will often contain a payload in the form of a malicious URL [23]. This form of contact is typically very successful, with [23] claiming a click-through rate of 66%. The compromise phase is where malware is installed on the target account's device, which is used to access the social media platform [23–25]. Finally, the contagion phase uses the compromised account to infect more accounts that are related to the target account [23].

Automated Spear Phishing is an attack whose effect can be exacerbated with information leaked from social media [26–28]. The attack relies upon the generation of *realistic emails* that targets will trust and will be tempted to follow the link, which will lead to either a compromised site that steals credentials or a payload that will compromise the target's machine. Recent developments in the field of Natural Language Processing (NLP) such as BERT [29], GPT-2 and GPT-3 [30], as well as XLNET [31] has improved the state of the art in Natural Language Generation (NLG) [32] and Natural Language Understanding[7] which will assist exploit tools by allowing them to generate more natural synthetic exploit emails which in turn will increase click-through rates on malicious emails.

3.2.2. Fraud

In UK law, fraud offences are governed by the Fraud Act of 2006 [33]. The Fraud Act defines several offences, including Fraud by False Representation, which arguably covers the offences discovered in the literature review, Astroturfing and Employment Scams [34].

3.2.2.1. Astroturfing. Astroturfing is "a fake grassroots activity on the Internet" [35] whose main aim is to influence lawmakers as well as elections and election campaigns. Astroturfing can be achieved using groups of bad actors or automated techniques such as social bots, or a combination of the two [36]. Astroturfing campaigns will use multiple sources to publish and promote the same message. This characteristic is likely to persuade the targets rather more than an Astroturfing message that came from one source [36]. Several techniques can be used to automatically detect Astroturfing campaigns on social media [37–42]. In the review of the literature, there were two main techniques discovered for detecting Astroturfing campaigns. They are author attribution and text classification.

Author attribution borrows techniques from Forensic Linguistics, where a single author is deduced for multiple texts. The problem is assumed to be analogous to plagiarism. This technique uses the multi-post characteristic of Astroturfing. This is a strategy proposed by [39]. They used a well-established technique from the Forensic Linguistics field of using word ngrams to identify similarities between texts, and therefore establish a common author. The drawback of this solution is that it may work well with manual Astroturfing campaigns which use a central script, its effectiveness against language generation techniques used by social bots is less certain because it is unlikely that social bots will replicate the same texts, but generate different texts on the same subject.

Text Classification is a technique that classifies texts from social media into Astroturf or a related term or Non-Astroturf category. The techniques discovered in the literature review are supervised techniques where a dataset was gathered from social media platforms and annotators labelled the data into appropriate categories. The annotators would be guided by a set of rules, for example, [40] stated that an Astroturf post is "a significant portion of the users involved ..... appeared to be spreading it in misleading ways" [40]. The authors produced a dataset of 366 documents, of which sixty-one were labelled as Astroturf. They compared an ensemble (AdaBoost) to a discriminative classifier, Support Vector Machine (SVM). The dataset is imbalanced, and consequently, the authors also used resampling techniques to balance the data set. The best results were achieved with AdaBoost and resampling that achieved an accuracy of 96.4%. However, the weakness of the paper is that the data set is small, and the number of Astroturf posts were only 16.66% of the data set. Therefore, the variety of the Astroturf posts in the dataset will be low, and this technique is unlikely to scale to classify the variety of Astroturfing posts found on social media.

In common with other crimes that are committed on social media, Astroturfing requires a form of Natural Language Generation, where manipulative social media posts are generated automatically. Natural Language Generation tasks are now typically undertaken by large Neural Networks [43], and this is the approach suggested by [42] for automated Astroturfing attacks. However, the authors claim that there are limitations to language generation techniques using Neural Networks, and these weaknesses can be exploited with a supervised classification technique. The main limitation is that during training Neural Networks lose information, and consequently, the generated posts will diverge from the source material [42] The loss described by [42] manifests itself in the form of grammatical patterns or sequences of words.

3.2.2.2. Employment Scams. Social media is used to commit Employment Scams. Employment Scams are offers of seeming legitimate employment but induce the victim into unwittingly committing crimes. Social media is used as a recruitment tool to find and target vulnerable individuals with false offers of employment. Representative Employment Scams that were found in the literature were money mules and reshipping.

Money mules are people who transfer money that has been acquired through illegal methods. Mules are often unaware that the money transfers are illegal, and they are paid a relatively small amount of money. Mules often believe that they have legitimate employment [44]. Mules need to be recruited, and social media can be a fertile ground for recruiters to find and recruit mules.

Reshipping scams recruit people to work from home where they are sent goods that have been bought with stolen credit cards, and they then forward it to the criminals who originally bought the merchandise. This scam makes it more difficult for the merchants to detect criminal behaviour because the goods are being sent to multiple legitimate addresses [45].

Recruitment for Employment Scams can be as simple as job offers on social media [46]. There are several indicators that an employment offer posted on social media is fraudulent. Although there is no formal academic research, there have been interviews with "scammers" who indicate that: "Posts that promise large amounts of money for very little work" [47] and "Employers that use the candidates own bank account to transfer their money" [47] are likely to be scams. In common with some criminal acts committed on social media, recruiters target users who have a specific profile, which in this case were individuals suffering financial strain [48].

The literature review failed in revealing techniques specifically to detect recruitment scams on social media, however, the literature review did find techniques that identified fraudulent job advertisements [49,50], and it seems reasonable to assume that these techniques can be used on social media. The discovered papers proposed one main approach for detecting Employment Scams, which is to use the job advertisement content [49,50]. The content approach is demonstrated by [49] who used contextual features from the advertisement

[7] https://archive.fo/i2dgq.

to infer that the advertisement is fraudulent. The contextual features they found that assisted in the classification are the existence of a website, the age of a website, and the existence of a LinkedIn page. These contextual features are supplemented by textual features such as Spam Words, and Structural Features. Structural features are features which describe employment requirements, interview process, employment benefits or company description [49]. Structural features include whether the employment advertisement contains job skills or describes remuneration [49]. These features are combined, and three classifiers were evaluated which were: a rule-based classifier (JRip), a decision tree (J48), and a probabilistic classifier (Naive Bayes). Each of these classifiers had a similar performance, with an accuracy of between 83.42% (Naive Bayes) and 96.15% (JRip).

The content approach is the dominant technique because the discovered papers were for job boards, however, it is quite clear from anecdotal evidence and small scale surveys [51] that job scams are prevalent on social media, and therefore techniques such as account profiling, and features which are used in similar crimes as defined by the Taxonomy in Fig. 2 are likely to be successful in job scam detection on social media.

3.2.3. Terrorism

In the United Kingdom, terrorist offences are defined by the Terrorism Act of 2000 [52]. The unique offence that was found in the literature review was Radicalisation. Radicalisation using social media is an offence under Section 6 and Section 58 of the act [52]. The act has an extraterritorial effect, and consequently, it is immaterial if the offences are committed outside the UK.

3.2.3.1. Radicalisation. There have been a number of definitions of radicalisation, and for the proposes of this survey, two definitions of radicalisation [53] will be followed: violent and non-violent radicalisation. Violent radicalisation is "a process, by which a person to an increasing extent accepts the use of undemocratic or violent means, including terrorism and nationalism, in an attempt to reach a specific political/ideological objective" [53]. Non-violent radicalisation is "the (active) pursuit of and/or support to far-reaching changes in society which may constitute a danger to (the continued existence of) the democratic legal order (aim), which may involve the use of undemocratic methods (means) that may harm the functioning of the democratic legal order (effect)" [53].

Radicalisation on social media is not a new phenomenon, as the RAND Corporation was reporting about radicalisation as far back as 2010 when the UK's Counter-Terrorism Internet Referral Unit was ordering take-downs of radicalisation material from social media [54]. Radicalisation is a process of recruitment that has several stages, which are: 1. Netting, 2. Funnel, 3. Infection and 4. Activation [55]. The Netting stage uses "narrowcasting or propaganda that targets specific sub-populations according to demographic factors (such as age or gender) as well as social injustice or economic circumstances" [55]. The Funnel stage is where candidate recruits are whittled down to several people who are likely to be easy to radicalise. In the infection stage, the remaining candidates are directed to resources and materials to "self-radicalised" [55]. The final stage involves the activation of the candidates to commit terrorist actions [55]. The use of social media typically concentrates upon Stages One and Three, as they need to use platforms such as Twitter to identify targets and disseminate propaganda material.

Some measures can estimate the risk of specific social media platforms to be used as a radicalisation tool [56]. The aforementioned article states that two types of metrics can be used to estimate the risk of a specific social media platform [56]. They are keywords and writing style [56]. The keyword approach looked for attitudes. The attitude keywords looked for "perception of discrimination" [56], nationalism, anti-western attitudes, and positive attitudes towards religious or nationalist extremists [56]. The writing style analysis identified personality types. The personality types susceptible to radicalisation are introverts who are frustrated with their current situation [56]. The approach described by the papers can identify users at risk who are likely to go through the radicalisation recruitment pipeline.

Accounts whose sole purpose is to radicalise vulnerable individuals will have a "tell", which allows the automation of their detection. In common with similar offences shown in the Crime Taxonomy, several approaches can detect an offending account. These approaches can be categorised as content, account profiling and information propagation. These techniques are likely to be successful because the raison d'être of the accounts will force them to behave and post in a specific manner. Various techniques have been proposed to detect such accounts [57–66].

A representative paper of the content-based approach is [57] who used hashtags such as #stealthjihad, #myjihad and #extremists to locate large collections of Tweets that are designed to radicalise the readers of the Tweet. The authors use an unspecified semi-supervised learning technique to gather more data. The training data is then used in the following one-class classification techniques: KNN and SVM, to classify unseen examples. The point of weaknesses for content-based techniques is that language used in radicalisation Tweets is likely to change depending upon the culture and education of the radicalisers as well as the natural evolution of the language, which is designed to avoid automated filters [67].

An extension to content-based techniques to estimate the emotion of the Tweet. As discussed earlier, radicalisers are looking for a specific type of person who can be affected by radicalisation Tweets. Emotion in Tweets can be used as a manipulation tool. This is the approach used by [58]. Their approach extracted emotion words from Tweets and represented them as a vector. In this approach, an Emotion Vector is computed for a domain, which in this case is radicalisation, and candidate vectors from unseen Tweets can be compared. Similar Emotion Vectors indicate that the Tweet it was drawn from is a radicalising Tweet.

In common with other crimes described in this section, some approaches use account profiling techniques. Accounts that are used for radicalisation will have a specific profile, which can be defined by features such as the number of followers and posts. This is the approach followed in [62] who used features such as "Number of posted tweets, Number of favourite tweets, the average number of hashtags, number of tweets per day and the interval between two consecutive tweets" [62].

Another factor in the radicalisation process is the influence of the radicaliser's social media account. Accounts that have weak, or no influence will not be successful in the radicalisation process. The removal of highly influential accounts will have a disproportionate effect on the radicalisation process. Influence in social media is a well-known problem and relies upon measures such as in-degree centrality or page rank where the influence is estimated by the in-links on a graph, which in social media would be followers. The problem of radicalisation influence is not just the number of followers an account has, but the successful propagation of radicalisation material (engagement), as well as the number of users pushed through the radicalisation recruitment pipeline.

There are some proposed approaches which detect the influence of accounts that are used in the radicalisation process [63,66,68,69]. For example, [68] used the content of a Tweet to predict its influence and by extension to the account it was posted from. The propagation of radicalisation information can be used to identify the account it originated from. This is the approach suggested by [66] who used Hidden Markov Models (HMM) to estimate the originating account of the radicalisation material.

3.2.4. Defamation

In the United Kingdom, defamatory statements are statements that have "caused or is likely to cause serious harm to the reputation of the claimant" [70] and are covered by the Defamation Act of 2013 [70]. Although the act limited the extraterritorial effect of defamation claims in the UK, it still has an extraterritorial effect in limited circumstances, as per the case of Soriano v Forensic News LLC [71]. Consequently, the location of the defendants in an action brought under the defamation act is immaterial, but the UK needs to be the natural location for the action. The unique offence found in the literature review was Fake News.

3.2.4.1. Fake News. Some detection techniques rely upon characteristics of fake news published on social media to mark them for removal. The main characteristics of fake news that can be exploited by automated methods are: "1. the false knowledge it carries, 2. its writing style, 3. its propagation patterns, and 4. the credibility of its creators and spreaders" [72].

Fake news can be spread via social media not only manually through troll farms [73] and retweet networks [73,74], but also through automated methods such as social bots [75]. The social bots function similarly as described in the Spear Phishing section where they automatically post material that will play to the prejudices of people who then will propagate the material to people within their network. In the light of the findings of [76] it is doubtful that these techniques have been very successful since only a very limited number of people have actually come into contact with fake news.

The strategy employed by social bots is in the early phase of fake news propagation. Social bots identify influential users whom they hope to influence to spread the fake news to their followers [75]. A large scale analysis by [75] found that social bots produced a relatively large amount of posts per week (100) of which thirty per cent would go viral, i.e numerous social media users will interact or view the fake story. It should be noted that this study was limited to Twitter.

Several strategies can be used to detect fake news posts on social media [77–84]. Automated fake news generation and their spread will depend upon Natural Language Generation techniques. As stated earlier, the function of social bots in the spread of fake news is to target influential social media users. There are therefore three factors: linguistic content, type of propagation and user type, to consider for automated fake news detection.

In common with other crimes committed, techniques that detect fake through post content are supervised learning techniques, where known examples are shown to a learner and a model is produced that identifies unlabelled candidates. In common with recent developments in text classification, attention networks [80], large language models [85] and large Neural Networks [86] have been used to classify social media posts as fake news. In addition to the textual content, images often accompany alarmist text, and in the approach proposed by [86] they used a variation of a Convolutional Neural Network that took text and images as inputs.

Fake news content can often be sparse and noisy, and the aforementioned text classification techniques may not be accurate enough to detect numerous fake news posts. The alternative is to classify the spreader of fake news and the propagation path of the post. This is the approach favoured by [84]. They produced an embedding of a target user by producing a vector that represents a candidate user's connections. By using these embeddings, it is possible to compute a community as well as similarity with other users. The assumption is that social bots will have similar embeddings and be members of the same community. The authors also used a Long Short-Term Memory Recurrent Neural Network (LSTM) to classify the propagation path of a social media post. Using the embeddings and the propagation path allows the classification of a social media post as fake news without examining the content of the post. Variations of this technique which use textual content of posts, type of connection and propagation have been used to find friends of a user who spread fake news [79].

3.3. Abuse

This technique is where the offender uses social media to abuse a targeted individual or group of individuals. The abuse technique may include threats, racial epithets or insults based upon a protected characteristic of the target group or individual.

3.3.1. Criminal Offences

Abusive behaviour on social media is covered by multiple legislation in the United Kingdom. The Crown Prosecution Service of the United Kingdom has provided a summary of the offences committed by abusive messages and behaviour on social media [87]. Offences can include[8]:

- "Making a threat to kill, contrary to section 16 Offences Against the Person Act 1861" [87]
- "Making a threat to commit criminal damage, contrary to section 2 Criminal Damage Act 1971" [87]
- "Harassment or stalking, contrary to sections 2, 2A, 4 or 4A Protection from Harassment Act 1997" [87]

There were two offences found in the literature review that would be covered by the abuse categorisation, they are Hate Speech and Cyberbullying. Hate speech is where minority groups are subject to derogatory and insulting language about their protected characteristics. Whereas Cyberbullying involves the use of bullying language against the target. This language can contain threats and insults against the target.

3.3.1.1. Hate Speech. In some jurisdictions, social media posts that are determined to be "offensive" can be considered a crime, and the author may face a prison sentence, censure or a fine. It should be noted that there are several jurisdictions, such as the USA, where free speech protections ensure that hate speech is not considered a crime. The term hate speech is derived from mature themes such as "flaming", hostile messaging and cyber-bullying [1]. Hate speech is a hard task for automated detection because hate speech is defined by the victim group, and therefore what could be considered legitimate criticism by one group could be interpreted by the target group of the criticism as hate speech. The recent issue of the Gender Critical community's criticisms of the Trans community [88] demonstrates the nebulous nature of hate speech.

There are several attempts to define hate speech, which then inform how hate speech material is collected and classified. For example, [89] described hate speech for their technique as "any offence motivated, in whole or in a part, by the offender's bias against an aspect of a group of people" [89]. Nevertheless, there have been a number of attempts to develop automated system that detect hate speech [90–95].

The work in [91] is a typical example of the approach used to detect hate speech on social media. The data that they used was collected from Twitter around what they called a trigger event. The trigger event in question was the murder of Private Lee Rigby by extremists in 2013. The data was collected for two weeks, which contained the hashtag #Woolwich, which is the location in London where Private Rigby was murdered. The rationale was for the two-week window was that public interest would peak and start to decline within that window. In total, they collected 450,000 Tweets. From these 450,000 Tweets, they selected 2000 Tweets for manual annotation. The paper is excluded how they defined what constituted hate speech, although they provided the following example: "Totally fed up with the way this country has turned into a haven for terrorists. Send them all back home.". They did not provide any inter-annotator agreement measures because they used a crowdsourced annotation platform. The features were a bag of words (most informative 2000 features) or hand-selected *hateful features*. The authors compared the results of a Bayesian logistic regression, Support

[8] A full summary of offences can be found at: https://archive.fo/uMkZc.

Vector Machine, Random Forest, and an ensemble of the three learners. In addition, they estimated the difference between the feature types. In this case, hateful features gave superior results to the most informative features.

The surveyed papers despite claiming to classify hate speech focus upon insults that are motivated by a poster's protected characteristics such as sex, race, and sexual orientation. This paper claims that this type of speech is a subsection of hate speech, and that a system that can classify all types of hate speech is beyond the capability of machine learning techniques because hate speech is defined by the target of the speech. Therefore, it is a subjective standard, and any dataset generated for this task will have a high degree of disagreement between the annotators.

3.3.1.2. Cyberbullying. Cyberbullying is arguably the antecedent of hate speech, where vulnerable individuals are bullied using social media and other Internet-enabled tools. The act of cyberbullying is defined as "an aggressive, intentional act or behaviour carried out by a group or an individual, using electronic forms of contact, repeatedly and over time, against a recipient who is unable to easily defend him/herself" [96]. Although cyberbullying may not include acts of violence, it does have consequences for the victim, which may include: "depression, low self-esteem, behavioural problems, and substance abuse" [96]. These real-world consequences of cyberbullying make the detection and removal of cyberbullying posts an obligation of social media companies. And in common with crimes discussed in this section, it is not possible to manually police all activity on social media platforms, consequently, an automatic approach is required.

There are relatively many approaches for the detection of cyberbullying on social media, and the following [97–103] is a representative sample of the techniques discovered. And again in common with other crimes in this section the three main approaches are: content [97–99,101–103], account profile [100] and a combination of the account and content features [104].

The content approaches are in the main two-class classification approaches where the learner is trained on a specific set of features or the learner is representational. Unlike the content approaches in other crimes, the sex of the poster plays an important role [99] because the type of language used by female cyberbullies is different to that of male cyberbullies [99]. The use of gender features can improve the performance of the base learner.

The content approaches can be classified into: automatic feature learning based on Deep Learning [97,98], Bullying Features [103], Rules/Decision Tree [102] and Sentiment [101]. The advantages that a deep learning model e.g. Convolutional Neural Networks (CNN) bring is that they are representational, and therefore no feature selection is required. Also, transfer learning [105] can be used to reduce the number of labelled examples required, and relationships can be learned from unlabelled corpora to pre-weight terms. A reproducibility study conducted by [98] which evaluated several techniques that used CNNs and their variants such as Recurrent Neural Networks (RNN) on three publicly available datasets [98] found that a form of RNN, bidirectional Long Short Term Memory (BiLSTM), outperformed traditional CNNs on detecting bullying posts. In addition, they evaluated several forms of transfer learning and found that model-based transfer, which transfers Neural Network Weights and Word Embedding from one domain to another, was the most effective in detecting bullying texts.

In common with other crime detection methods, in this section, account features such as the number of followers, posts and so on can assist in the detection of cyberbullying posts. Sock puppet and impersonated accounts on social media can be used to bully people on social media without revealing their true identity, consequently, user features can improve the detection of bullying posts [100]. This is the approach taken by [100] whose main user feature 'age' was used to improve detection of bullying posts. Finally, [104], uses combinations of account features, and content of posts to deduce if a post is a bullying post.

3.4. Learner selection

This section will describe the popularity of the learners used in the detection of crimes on social media. This is because the accuracy of content-based approaches will be in part be determined by not only the data selection technique, but by the choice of learner (Classifier). The choice of the learner should be determined by the data rather than by any predetermined bias. The majority of the papers surveyed in this area did not have a learner selection description, or at least a rationale for their learner choice. Nevertheless, this section will show the learners used in the aforementioned crime detection tasks and their percentage use in the referenced papers. Learners that are referenced once are grouped under others.

These values are shown in Fig. 3, and it is quite clear that the single most used learner is the Support Vector Machine (SVM). SVMs are used in older papers or are used as a baseline to compare to more modern techniques. Researchers who work in this domain seem to be following the general adoption of large Neural Networks in the mainstream NLP Community. New research in this area will likely follow this trend.

3.5. Representation of data for supervised learning

In common with learners, how input data is represented as features to a learner and classified by a model will have an effect upon the accuracy of the technique. Representation learners such as Neural Networks learns features automatically, however they can rely upon an embedding layer that represents the vocabulary in texts as a vector. This vector captures the semantic properties of a term or word. The embedding technique had a great impact upon the model's accuracy [6].

The main embeddings used are Aspect [80], which is a variation of word embeddings where the vectors represent an aspect, a property of an object, and their co-occurrence with terms. Comment Embeddings [95] is where posts are represented as paragraph vectors which can represent variable-length text as variables, Context-Aware Embeddings [85], which are embeddings that represent the different contexts of the same words, Sentiment-specific word embedding [78,97,98], which are embeddings that represent sentiment words, Word Embedding [81,86,97,98], which are embeddings that represent words in a Tweet/Post, User Embeddings [83,84], which encodes account information about a user on social media, and News Embeddings [83] which encodes news information from a Tweet into a vector.

The older learners require Feature Selection techniques to reduce the number of features and consequently the amount of data required to produce an effective model. The survey revealed several frequent approaches. The majority are based upon the content of a post, however, several approaches rely upon features of the poster themselves. The main content approaches use the following features:

- Hateful Terms [92], which is a subset of words or phrases that have been pre-determined to be hateful,
- Typed Dependencies [92], which are words that are grammatically linked
- News Features [79,82], are features drawn from news stories and represent: headline, source of news, news content, and features from visual elements
- Part of Speech (POS) [42,75,90,91,101,102] which are features derived from POS taggers
- Term Frequency Inverse Document Frequency (TF-IDF) [38,99,101,102] are keywords which are identified using
- Profane Words, are words or phrases that are generally considered to be obscene
- Term Frequency [59] is the number of times a word occurs in a collection of Tweets/Posts
- Document Frequency [59] which is the number of individual posts a word occurs in a collection of Tweets/posts

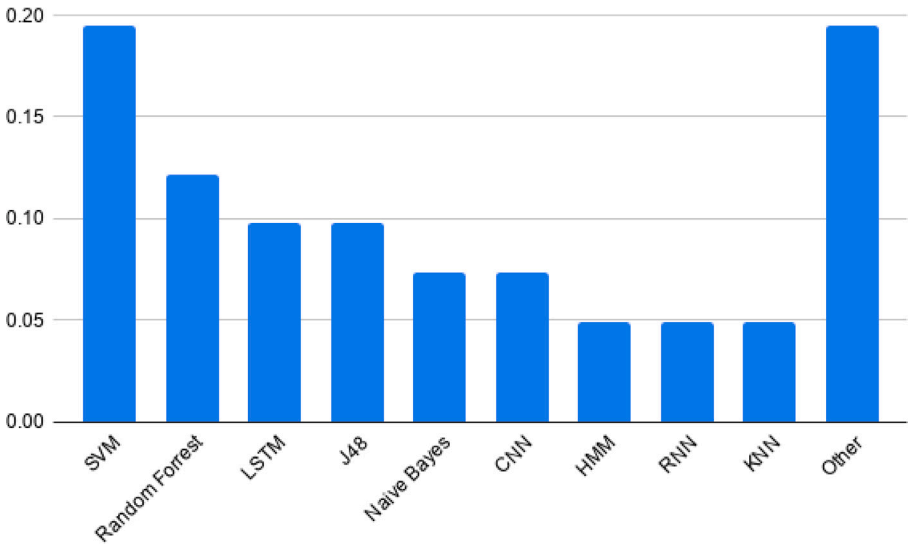


Fig. 3. Learner use by percentage.

- Sentiment [22,42,75] which are words and phrases that carry subjectivity information
- NGrams [39,99] which are Multi Word Expressions that contain two Words or more
- HashTags [40] which is the extraction of HashTags from posts
- Names [40] which are names of people in the posts.

The main account features that were discovered in the literature review are

- Gender [101], which is an indicator if poster is male or female
- User Features [22,41,58,62,75,81,99,100,104], which are features derived from the account of the poster such as account creation date and username
- Temporal Features [22,41,58,62,75,77,104], which are timing features related to number of posts per day, timings between posts and so on
- User Frequency [59], which is the number of individuals who use the word
- News User Features [79,104] which are features from the user's network.

Two approaches could not be classified as content or account techniques. These features are Visual Clues [82] which are features from elements such as videos and images, and Structural Features [37,42,50,104] which are meta-features about a post such as the number of words and sentences.

The most frequent approaches when selecting embeddings or features for content-based techniques are ones that carry opinionated information such as profane language, hateful terms and sentiment features. This should be unsurprising because many of the crimes described in this section describe intimidation or manipulation of targeted individuals, which require subjective and opinionated language.

The most frequent account features and embeddings use temporal and users features. This is due to techniques designed to discover sock puppet accounts, which have a specific account profile, and automated posting which is likely to be conducted on schedule. And therefore can be detected through the use of temporal features.

3.6. Common approaches

This paper has a hypothesis that crimes that have similar exploit methodologies will have common detection methods. This subsection will discuss the exploit methods of manipulation and abuse, as these methods are the ones that have two or more crimes grouped under them.

3.6.1. Manipulation

The main approach for detecting manipulation offences is text classification [40,49,57], and it is an explicit suggestion of this paper that text classification would be a suitable technique for detecting exploit methods that use manipulation to induce an action or obtain information from a target. This is because the technique is often used at scale and certain trigger words and phrases are likely to have a high frequency within the exploit messages. In addition to trigger phrases, manipulative exploit messages are likely to have a high concentration of sentimental or emotional content, as these features can be used to manipulate the target [106].

3.6.2. Abuse

In common with the manipulation exploit method, the abuse exploit method has the common detection method of text classification [91,103]. The text classification approaches are successful within this domain because the abuse exploit method seeks to insult and demean the target. Therefore, this exploit method will contain an excess of emotional and insulting language, which if detected will be able to identify abusive messages.

3.7. Adaption to new crimes

Social media is a dynamic environment where new crimes and frauds are invented and committed, and it is not possible for the research community to generate new datasets and models for each specific crime. It may be possible to use suitable techniques from similar offences or reuse models and data from similar domains. For example, there have been reports in the traditional media that social media is being used for rent, healthcare, cancel and various other scams.[9] It was not possible to locate papers in the literature search that addressed each of these scam types individually. However, it is a hypothesis of this paper that these scams could be addressed with existing models or datasets and techniques such as Zero-Shot Learning [107] and domain adaption [108] to reduce the need for labelled data for a new crime or exploit.

[9] https://archive.fo/NxK3n.

3.8. Discussion

Social media is being used to commit crimes, and Social Media Companies have an obligation to stop and impede criminal behaviour. Social media companies cannot remove all criminal behaviour from their networks unless they use draconian user identification methods. This approach is unlikely to succeed as it will stifle the number of the users. Companies that have a liberal user identification method will surpass them because of fewer barriers to being a user. The alternate strategy is post sign-up intervention methods that identify criminal behaviour, remove the post, and ban the user. As shown by this section, there are several automated techniques that can be used to detect and ban offending users at scale.

A possible open issue is the transfer of data and techniques from one established crime to a novel crime. In this way, security researchers will not have a lag from a novel crime being detected and a solution being developed. A possible way forward is to a high-level detection where criminal behaviour is detected in a user's behaviour and posting rather than detecting specific crimes, and that techniques such as Zero-Shot Learning [109] and Transfer Learning [110] may help because they allow the reuse of data and models to new domains. Zero-Shot learning is a technique where a learner needs to predict a class for a data instance that does not exist in the training data [109]. An example would be a learner trained on horse images, detecting ponies in images in the test data. Zero-shot learning would do this by finding an association during the training phase. Zero-shot learning has been used successfully in machine learning with social media data [111], and therefore it is reasonable to assume that this technique will be successful on related crimes where there is an abundance of labelled data in one domain and a lack of labelled data in another. Transfer learning is a series of techniques that are used "to improve a learner from one domain by transferring information from a related domain" [112]. Transfer learning has been used successful in training a model for Hate Speech [113], and therefore it is reasonable to assume that this technique can be expanded to other related crimes.

4. Exploit methods

Social media cannot only be used to commit crimes, it can also be used to facilitate crime by offering an attack vector. Information can be gained and targets induced to visit malicious sites where their system will be compromised. This section will discuss several exploit methods using social media, such as Information Leakage and Impersonation.

The motivation of this section is to survey methods that are used by criminals to gather information from targets to exploit the target or systems that they control access to, such as bank accounts

4.1. Information leakage

Information leakage is the process where protected information appears on social media through the negligence of an individual, company, or organisation. The type of Tweets that expose possible information which can be applied to various crimes is summarised by [114]. The summary contains the types of social media posts that can disclose information as Vacation Posts, Drunk Posts, and Disease Posts. Vacation posts often reveal information that the user is on holiday, and with other information, this may allow criminals to target the user's house for various crimes. Drunk posts are where the poster is inebriated and may reveal personal information such as sexuality or the committing of crimes. Disease posts reveal information about the poster's health.

An obvious form of information leakage is where users publish their personal information such as physical location, name, and job role. This information can be used to enable crimes where attackers use personal details to steal identities.[10] The victim of information leakage may often not post their own details to social media because friends or relatives can annotate posts, and images with personal information such as real name or date of birth [115].

The nature and impact of threats of social media were described by [116] and are shown in Table 1. They identified four main areas where information can be leaked from social media.

Information leakage is not limited to personal information, and users may unwittingly leak information that can be used to exploit the user's employer. The typical type of attack using this information is social engineering. Social engineering [117] is a psychological attack where an attacker gains the confidence of a gatekeeper to a secure system and uses that trust to exploit the secure system. For example, CEO impersonation [118], could be a more credible attack when the attacker has access to time-sensitive information such as vacation information. In this hypothetical example, the attacker could use the area code in a spoofed telephone number when contacting the target company.

Information leakage from social networks can assist a specific attack known as a CEO impersonation attack. The CEO impersonation attack is an example of using social engineering to make specific attacks against high-value targets, e.g. an incident that happened in Ireland county council which resulted in a transfer of 4.3 million euros to an account in Hong Kong [7]. The move from random attacks to specific attacks has become a more popular form of attack [119]. Information leakage can assist in more specific forms of attacks, such as "Spear Phishing". Information leaked from social media can improve phishing attacks by generating emails that contain content that targets will click [116]. The combination of information leakage and language models are likely to improve the effectiveness of automated Spear Phishing.

4.2. Estimating social media risk

Social media users may be ignorant of their exposure to information leakage. There are, however, techniques to estimate the information leakage across multiple social media platforms [120]. For example, [120] used an aggregate of information voluntarily released on a number of social media platforms to estimate the average number of attributes for several common attacks. The common theme of these metrics is that the attack vector grows with the number of social profiles that the individual creates [121]. Information leaked from social media can be aggregated with non-social information such as phone directories to create a greater attack vector [121]. These metrics can be used to predict the likelihood of insider attacks on networked computer systems [122].

The aggregation of multiple accounts created by the same person on different social media platforms is not a trivial task, there are however a number of techniques [123–127] that can be used to achieve this task. A representative example of matching profiles on multiple social media platforms is provided by [127] who used a similarity-based approach to match users on the LinkedIn and Twitter platforms. The similarity matching uses publicly available profiles to match parts of the profiles such as name, location, interests, and so on. Text comparisons were achieved using Jaro–Winkler distance, which is a string similarity measure. Geographical comparisons initially normalised locations using Geonames ontology. The normalised approach computed a geographical bounding box, for example, New York's bounding box would be the city itself. The similarity between the two locations would be the Euclidean Distance between the centres of the bounding boxes of the different locations. In this way the similarity of a city, e.g. New York, with the sub-area of a city, e.g. Queens could be computed. The educational similarity is computed using Smith–Waterman distance, and the similarity between summaries was achieved through SoftTFIDF. These similarities were then used as features in various classifiers. This

[10] https://archive.fo/qhq6c.

Table 1
Summary of information leakage [116].

Action	Security threat	Impact to organisation
Status update	Status update information is accessible to everyone on the social network, from which sensitive information may be revealed.	Confidential information can be obtained from the status updates.
Friend requests	Unfiltered accepting of friend requests can result in fraudsters/attackers being accepted.	Friends have access to more information than other users of the social network. It is easier for "friends" to gather information for an attack.
Photos and videos	Careless posting of images can reveal sensitive information	Information from images can allow attackers to gain an insight on how to compromise a person or organisation's system
3rd party apps	Apps can be gateway for malware which can compromise the user' computer or phone	Compromised system can be used to access the organisation's system

approach in its current form is not suitable for large scale matching because in the worse case to gain one profile match the technique will need to search n profiles where n is the number of profiles on a social media platform, which is for Twitter 330 million.[11] Besides, the similarity measures are computationally expensive, which would prohibit any large scale exploit. Despite the limitations of content-based profile matching, it does show the risk of matching publicly available information.

4.3. Impersonation

Automated fraud perpetrated on social media can often rely upon impersonated accounts. Impersonation is the act of creating a social media profile or page which purports to represent a legitimate person or company. The impersonation of an organisation may include techniques such as small imperceptible spelling changes of a company name and in the case of individuals the unauthorised use of pictures and their name. Impersonation has been used to commit frauds such as romance scams [128], cyberstalking [129], and as previously discussed, phishing.

Social media sites for well-known brands and individuals have a form of authentication, however social media sites that do not verify users' identity are prone to a type of attack known as a Sybil attack [130]. A Sybil attack is where an impersonated account acts as an *honest broker* to directly target users to compromised websites or other locations, where a payload such as malware is used to infect the target's account.

Impersonated accounts can be used to communicate manually created messages, however, to comprise accounts at scale require automated methods. Automated methods for impersonation will use a bot to generate posts that will be used to commit further crimes, such as the aforementioned spear-phishing and financial crimes [131]. The role of automated posts in impersonation of legitimate social media accounts has already been discussed, as well as the detection of fake or impersonated profiles on social media. However, despite the interest of social media companies in detecting and removing impersonated accounts [132], impersonation will continue to be a popular method through which malicious actors can commit and encourage crimes.

4.4. Discussion

Information leakage from employees can increase the attack vector against an organisation's infrastructure, this risk can be mitigated by the implementation of corporate social media use policies as well as logging and monitoring of users' computer use [133]. These policies only manage the use of corporate computing resources and are unlikely to curb information leaks from social media posts made from private computing resources. Information leakage from social media will be an area that attackers will exploit for the foreseeable future.

An open problem in this area is to detect information leakage as it occurs, or to identify users who are likely to unwittingly leak information and halt the leakage as it occurs. Retroactive measures are likely to fail because the information has already been propagated onto social media and may be used by attackers.

5. Prevention and detection of crime

Social media is now part of a crime agency's arsenal to detect and predict crime. The crimes may be trivial such as theft or serious such as violent disorder and public insurrection. The predictors of crime may be threats against a specific person or organisation or information gleaned from multiple posts. This information can be used by law enforcement to prevent crimes.

The motivation of this section is to identify methods that are used to resolve or report crimes that are committed in the physical world from information on social media.

Social media has changed how criminals act after they have committed a crime. Traditionally criminals would want to hide evidence of a crime, but since the advent of social media, criminals have started to publicise their criminal act [134].[12] This information can be used to not only identify a criminal, but can be included in the evidence in some jurisdictions at a criminal trial. This section will cover work in the areas of crime prediction, criminal detection, and detection of impersonated accounts.

5.1. Crime prediction

Social media can contain information that allows the prediction of a specific crime or a group of crimes against specific targets or within a certain region. Information from social media can be used discreetly or with other information sources. This area of research is a relatively popular area where there are a number of papers. The articles found in the literature review are [135–147]. The main areas addressed by the research are the prediction of Distributed Denial of Service Attacks (DDOS) [136,145], Civil Unrest [137–139,148–150], Hit and Run [147], General Crime [141,143,146], and Robbery [142].

The prediction of DDOS attacks against organisations is typically predicted by the sentiment directed against a particular organisation or

[11] https://bit.ly/3g5CK10.

[12] See https://archive.fo/quRLg as an example.

entity through social media messaging. Negative sentiment directed towards an organisation is a strong predictor of future attacks [136,145]. The assumption is made that sentiment on social media reflects general public opinion [151], and that there will be a minority of people who are motivated by the negative sentiment who will take direct action [152]. The anger towards certain entities generated through social media also can have real-world consequences and this hypothesis seems to reflect the motivation of criminals involved in the publication of fake news, where false negative sentiment against an entity is intended to influence voters' behaviour. The underlying rationale for these techniques is that emotion drives behaviour.

Civil unrest has been a common theme in research and has been applied to popular uprisings such as the Arab Spring [149]. The hypothesis behind using social media to predict civil unrest is that groups often use social media to organise future protests[13] and that these messages hold information that can be used to predict the location and time of the protest or the violent disorder [137–139]. The main approaches found in the literature review for predicting civil unrest are Information Propagation [148], Activity Cascades [139], Text Classification [137, 149,150,153].

Information propagation in social networks is where information passes from one to one or more users. In [148], an example of information propagation is given as: "Twitter user Alice posts a tweet1 on a protest event on a given day. Bob, a follower of Alice also posts a tweet on the same protest event after the original post by Alice" [148]. The technique developed by [148] uses the characteristics of information propagation of Tweets that contain protest information to estimate if the protest will take place. The authors describe information propagation as a tree and the features from these trees were used to train a Support Vector Machine (SVM) which was used to predict the likelihood of a protest occurring.

Activity cascades could be seen as similar to information propagation, but in the approach described by [139] they claim that activity cascades are where a post by one user causes subsequent posts on a similar subject by other users. They claim that large activity cascades are indicators of large or important events. They developed a regression model from features derived from activity cascades to predict future civil unrest.

Text classification approaches look for specific information in a post or tweet which indicates civil unrest at a future date. Text classification approaches can use unsupervised methods such as clustering [137,150] or supervised techniques [149,153]. Supervised techniques can be augmented with semi-supervised strategies such as Active Learning [153] to improve the accuracy of the model.

Hit-and-run offences are where a driver has a crash with another vehicle or person and leaves the scene of the accident. This is a crime in many legal jurisdictions. The unique paper found for this area [147] followed text classification approaches used in the civil unrest domain, where specific information related to offences were extracted from Tweets. The authors used Topic Modelling to make links between words in Tweets and hit-and-run incidents. The topic distribution is used in a Generalised Linear Model to predict hit-and-run offences from Tweet information.

General crime for this paper is where the offences being committed span more than one type of crime. For example, the technique described in [143] detected twenty-five crime types. The hypothesis of crime predictors in these techniques is similar to that of other domains where words in a social media post function either as a statement of intent, or predictors of crime. However, there was one exception to this hypothesis which was [146] who assumed that social media posts can be used to estimate the concentration of people at specific places, and the concentration of people is an indicator of future crime at that location.

The robbery predictor paper [142] is similar to approaches in other domains such as hit-and-run, and general crime, where specific information, in this case car descriptions, is extracted from Tweets to use to predict robberies.

The papers discovered in this section can reflect the type of crime they are trying to predict. Crimes that require mass participation rely upon the inciting information propagating across social media platforms. Whereas crimes against organisations can be predicted by detecting the wider reputation of the organisations. These can be found using techniques such as sentiment or emotion analysis. Finally, specific crimes committed by individuals can be inferred by crime-related words or topics are found in social media posts. These techniques should be seen as a complement to proactively policing techniques [154], not as a replacement.

5.2. Detection of criminals

Information on social media can be used to detect criminals and criminal activity. Information leakage not only can be used as an exploit vector, but also can be used to identify criminal activities and networks. Detection of criminals on Twitter can be a simple task because criminals can often boast about their crimes. For example, the person who undertook the CapitalOne exploits boasted about her crime on various social networks.[14] Not all criminals give such obvious signals, and therefore more sophisticated techniques are required to detect individual criminals or networks of criminals. Some techniques can detect subtle signals and imply that the poster is a criminal or part of a network of criminals [155–162].

A typical example of the techniques used to identify criminals is provided by [160]. Their approach is a two class classification approach that divides posts into a criminal or non-criminal category. These posts then allow the identification of criminal networks. The approach is essentially a content-based approach that uses topic modelling. The paper does not however provide any experimental evidence that the technique is successful in identifying criminal networks, nor does it provide any data of what a criminal tweet or post consists of. Despite this, the technique is a simple explanation of the content-based analysis of social media posts for admissions of criminal acts. Although, it is unlikely that this approach can be used on a large scale because the technique would need to scan all posts on a social network.

As argued earlier, content-based social network analysis has a point of weakness, and therefore the techniques are unlikely to scale, and it is an issue that the majority of authors ignore. There is one exception, i.e. [162], which describes a system that is intended to process numerous irrelevant social media posts. Their technique relied upon software such as Apache Spark, and Apache Kafka, which is designed to process large amounts of data. However, it is arguable that even with High-Performance Computing (HPC) content-based approaches will not be able to process sufficient amounts of data to find sufficient numbers of criminals and their related networks. Content-based techniques need to have scalability as part of their design, where numerous social media posts can be ignored.

An alternate approach to a content-based to infer a criminal network on social networks is network analysis. The survey conducted for this paper did not reveal any research that explicitly used network analysis on social networks. However, there is related research that used other sources of information and network analysis to infer criminal networks. For example, [163] who used weblogs and their hyperlinks, as well as network analysis to infer criminal networks. It, therefore, seems reasonable to suggest the same technique but using social networks and publicly available friend lists as well as linked content in posts to infer criminal networks is feasible. This approach is likely to be scalable because it is likely that a seed set of known criminals will be used to select accounts that are likely to be part of a criminal network. This will limit the number of posts and accounts that need to be analysed.

[13] https://archive.fo/I4tEd.

[14] https://archive.fo/UJ25n.

5.3. Detection of impersonated accounts

Impersonated accounts are often the basis or enablers of crimes such as Spear Phishing. The automatic detection of impersonated accounts at scale can assist social media companies to reduce and control the amount of crime committed on their networks. A simple solution to reduce the number of impersonated accounts would be to demand some form of government-issued identification to be submitted before an account can be opened. However, this type of restriction is likely to impede the number of users and would limit the ability of the human rights activists to raise awareness of abuses in totalitarian regimes, as well as impeding whistleblowers from publicising actions of companies [164]. In addition, social media companies' operations span numerous legal jurisdictions with differing political orientations. It is highly unlikely that a world-encompassing social media identification policy could be agreed to. To circumvent national ID laws, criminals would simply spoof IPs of countries with lax ID laws to open a fraudulent account.

The alternative to a legislative approach is to detect characteristics of an impersonated account and either demand further identification of the poster's identity or delete the account. This approach can be used for user reporting. A simple method for detecting impersonated accounts is to look at the friends–followers ratio and the evolution of friends–followers ratio [165]. The friends–follower ratio for impersonated accounts is around thirty, whereas for legitimate accounts it is about one. Impersonated accounts tend not to gain friends over time, whereas legitimate ones will gain friends. A complementary approach to account profiling is the characteristics of the posts from the impersonated accounts. As demonstrated in the phishing sections of this review, impersonated or robot accounts posts will have specific characteristics that can be detected by automated techniques [166]. Impersonated accounts are not limited to high profile accounts such as celebrities and politicians. Normal and low profile accounts are just as likely to be impersonated as high profile ones [167].

There are several papers that describe detection techniques. [166–172]. The main approaches can be divided into the aforementioned categories of Account Profiling and Textual Characteristics, as well as the combination of the two approaches. Account profiling is a technique that uses the characteristics of the account to infer if the account has been impersonated. Typically, these techniques do not use the linguistic characteristics of the post. For example, [168] used "Education and Work, Gender, Relationship Status, Number of wall posts by the person, Number of photos of a person tagged in, Number of photos the person has uploaded, Number of tags in the uploaded photos by the person and Tagging Average on photos" [168] to deduce if the account has been impersonated.

Text Analysis is where the technique identifies some form of writing or posting style of the impersonated account. This characteristic may be a weakness in automated text generation methods or the raison d'être of the impersonated account. The text analysis approach was used by [166] who used an analysis of frequent and infrequent domain-specific words to define an account signature. This approach can detect: "authors with sockpuppets accounts" and "front-user accounts which are operated by several authors" [166].

The outlier in the survey is [167] which identified pairs of accounts that are "doppelgangers", where one account is likely to be an impersonated account of another. Their approach used a rule-based approach that detected similarities between *user-name, screen-name, location, photo, and biography* [167]. Their rule-based approach detects similarities between these attributes and makes a connection between the accounts. They also used the now-defunct Klout Score service[15] to estimate the reputation score of the accounts. With the reputation score, it was possible to determine from the doppelgänger pair which account is fraudulent and which is genuine.

5.4. Discussion

Social media not only offers opportunities to commit a crime, but also offers an alternative route for crime detection. Criminals and their crimes leave indicators of their future crimes, as well as their criminal associates. Automated systems that trawl social media can offer law enforcement a rapid way of identifying criminals as well as the location of future crimes. Social media will not replace traditional investigation techniques, but will complement them. There should be a caveat to the described techniques. Any technique that is successful at scale will provoke a change in criminal behaviour on social media to circumvent the technique. Criminal detection techniques will need to evolve to match criminal behaviour, and consequently there will always be a window of opportunity for criminals to exploit while automated systems catch up. And therefore, an open problem is to transfer detection techniques from one domain to another. As with crimes on social media, there will be a similarity between crimes committed in the physical world, and consequently it may be possible to transfer techniques between similar criminal offences.

6. Social media tools and data

The exploitation of information on social media requires some technical sophistication, which may be beyond the ability of attackers and security researchers. The availability of tools will allow low-skill or "script kiddies" to launch attacks using social media information. The motivation of this section is to document the available datasets and tools that security researchers can use and adapt to their research.

All the tools found in the review conducted for this paper were for Spear Phishing. There are likely more tools available, but because of the criminal intent of these tools, they are not published in the academic literature. A summary of a sample of the available tools is shown in Table 2. A representative tool of the ones found in the literature review is SNAP_R which uses a form of a Recurrent Neural Network (RNN) to generate phishing posts with embedded malicious links that a target will click which will execute a payload on the target's system. The system will then be exploited by the attackers. An example of posts generated by the system can be found in Table 3. As shown by the examples that the phishing posts seem "natural", and it seems reasonable that targets may click on the links. The release of large language models such as GPT-2 or more recently GPT-3 [30], as well as the increase of multiple social media accounts held by each individual, is likely to improve the click-through rates of tools such as SNAP_R.

6.1. Social media datasets

The papers discovered in the literature review for this paper rarely published their datasets, which means that it is often not possible to replicate the reported results. Replication is fundamental in any form of science, and the community should use freely available datasets. Therefore, in this subsection is a list of freely available datasets, and where available a link to the dataset is provided. The datasets are listed in Table 4.

Although academic researchers did not release their datasets, competitive machine learning platforms such as Kaggle which are used to host machine learning competitions have released numerous datasets to the public. These datasets can be novel and collected for the platform, or a combination of existing datasets. This is reflected in the datasets found in the literature review for this survey, where many datasets are only hosted on these platforms and are absent from the academic literature. Organised shared tasks are an alternative to commercial competitions, and they are perhaps more relevant to academic research. One of the more well-known organisers is SemEval which is part of the International Workshop on Semantic Evaluation.[16] SemEval provided

[15] https://archive.fo/8JQ6n.

[16] https://archive.fo/ci20r.

Table 2
Social media exploit tools.

Tool name	Reference	Description
SNAP_R	https://archive.fo/68oGx	The tool is used to automatically generate phishing posts for target Twitter Accounts. The tool uses an LSTM trained on a general corpus and a Markov model trained on the target account's public posts
Speed phishing framework	https://archive.fo/fCu1H	SPF was designed to automate Spear Phishing attacks by generating automated emails to targets
recon-ng	https://archive.fo/aiIBg	Recon-ng is a reconnaissance framework that is designed to identify possible targets for automated spear phishing attacks.
Vase (Vulnerability Analysis and Scoring Engine) [173]	No publicly available code	Vase uses Twitter discussions about Common Vulnerability and Exposure (CVEs) to predict e Common Vulnerability Scoring System (CVSS) scores before the official assessments from NIST.

Table 3
Example Tweets of SNAP_R.

Example Tweets[a]
@andrewmcgill Someone should do a story on the inequality of happiness within countries. https://t.co/HWENeSfs92
@kavehwaddell out of sight, out of sight, out of sight, out of mind. also, harder to tap. https://t.co/eKY0heVcoQ
@marinakoren Russia and Ukraine are nearing a deal for the Quiet Car on This Trump Train https://t.co/eKY0heVcoQ
@ibogost Welcome to my hood. This is great tho https://t.co/ZuDpl23qy7

[a] https://www.theatlantic.com/technology/archive/2016/08/the-twitter-bot-that-sounds-just-like-me/496340/.

one of the datasets, HateEval, found in the literature review. Finally, media organisations often release social media that they think are in the public interest. In this case, NBC a media organisation from the USA released a cache of Election Tweets from alleged *Russian Troll Farms*.

6.1.1. Cybercrime data repositories

AZsecure[17] data is a data science testbed for the intelligence and security research community, maintained by the University of Arizona's Artificial Intelligence Laboratory, AZSecure-data.org provides a list of social web data sources.. The site maintains its own Dark Web and Geo Web forum data collection,[18] and datasets of hate speech, as well as links to data maintained by other universities, such as the security-related Twitter data maintained by the University of Virginia, similar to the Outlier Detection DataSets (ODDS). While many of these datasets are related to network traffic and opinion fraud in online reviews, the list also contains a Twitter dataset from 2014 related to terrorism and domestic security. During the literature review, the authors have collected a list of datasets, API's, and platforms that can be used for detecting cybercrime. The majority of the datasets were collected from several websites. There are ten freely available network-related cybersecurity datasets, three phishing related datasets, one dataset is also available related to the detection of credit card fraud. Twitter has four freely available datasets, including those related to influence bots, spam, and terrorism.

6.2. Discussion

Exploring the techniques used by black hat hackers and bad state actors to exploit social media for criminal gain requires tools and datasets, both of which are in short supply. Researchers in the area as a rule do not release either the code, tool, or data that they use in their research. This is regrettable that academic researchers do not release their data and code, and the release of data and code must be mandated by academic journals and conferences so that results can be replicated by independent researchers. Without replication, there will always be some doubt on the results produced described in this survey. Papers in Computer vision and machine learning conferences like CVPR and ICCV release their code, consequently computer vision has rapidly improved. Similarly, there have been some initiatives from conferences such as ECML[19] to encourage reproducibility, and for research that has a societal impact, data submission must be mandatory.

In common with other discussion sections, this paper suggests that the ability to transfer or re-purposing existing tools and data to a new domain is an open problem, and is a method of adapting to new criminal behaviour.

7. Conclusion

This survey has demonstrated that social media is being used as an attack vector to commit criminal acts such as comprising information systems, fraud, and inducing users into committing terrorist acts. Social media companies have to police their sites to identify, remove, and ban users who are committing criminal acts. This is a delicate balancing act of freedom of speech and criminal intent. And as shown by the hate speech section, criminal intent is not always clear. Criminal acts such as Lèse Majesté[20] are often used to suppress criticism of leaders and which is legitimate behaviour in many jurisdictions. However, social media should not be used as a forum of hate speech in the name of freedom of expression. Social media companies must only consider criminal behaviours which are a breach of natural law, where the consequence of the post is a crime in most legal jurisdictions. The crime detection methods are similar for each type of crime, either they rely upon content, account profiling, and in some cases information propagation. The content-based methods that use large Neural Networks can be adapted for new areas using transfer learning, consequently, this paper suggests that researchers share their models via a central repository

[17] AZSecure-data.org.

[18] See related publication in the section on the underground economy.

[19] https://archive.fo/uE4mJ.

[20] For example, see https://archive.fo/xahej.

Table 4
Social media datasets.

Data set types	References	Location
Civil unrest		https://archive.fo/A0Gn2
Civil unrest		https://archive.fo/j9QWd
Civil unrest		https://archive.fo/HQk7O
Impersonation	[174]	https://archive.fo/5d6e3 https://archive.fo/KWyGN
Fake news	[175]	https://archive.fo/14lJS
Fake news		https://archive.fo/qNuWa
Hate speech		https://archive.fo/pTjI6
Hate speech		https://archive.fo/CKVgF
Hate speech	[176]	https://archive.fo/qahkm
Hate speech	[177]	https://archive.fo/6jWD5
Hate speech	[178]	https://archive.fo/9AptN
Offensive communication	[179]	https://archive.fo/H0eQG
Russian troll Tweets		http://nodeassets.nbcnews.com/russian-twitter-trolls/tweets.csv
Radicalisation		https://archive.fo/wctbB Kaggle
Employment Scams	[180]	https://archive.fo/QKYk7
Cyberbullying		https://archive.fo/bjtuD
Cyberbullying	[98]	https://archive.fo/qlWwk

such as a model zoo[21] so that researchers do not have to train Neural Networks from scratch because this can take large amounts of data. Transfer learning will reduce the amount of data required to produce an effective model.

The majority of the approaches and datasets found are for English, which limits researchers' who are working in a weakly supported language ability to research social media posts in their language. An alternative to building datasets in specific languages is to use multilingual frameworks such as Multifit [181] that allows training of a model in a well-supported language such as English, which will also be able to classify texts in a weakly supported language. This approach frees the researcher from language constraints. It is a strong suggestion of this paper that future approaches work in a language-neutral manner. It is hoped that these suggestions for future research areas will improve the state of the art and allow the adoption of the described techniques to niche crimes and languages.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Anna Schmidt, Michael Wiegand, A survey on hate speech detection using natural language processing, in: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, 2017, pp. 1–10.

[2] Ahmed Aleroud, Lina Zhou, Phishing environments, techniques, and countermeasures: A survey, Comput. Secur. 68 (2017) 160–196.

[3] James P. Walsh, Christopher O'Connor, Social media and policing: A review of recent research, Sociol. Compass 13 (1) (2019) e12648.

[4] Raja Ashok Bolla, Crime Pattern Detection Using Online Social Media (Master's thesis), Missouri University of Science and Technology, 2014.

[5] Alina Ristea, Chad Langford, Michael Leitner, Relationships between crime and Twitter activity around stadiums, in: Geoinformatics, 2017 25th International Conference on, IEEE, 2017, pp. 1–5.

[6] Ihsan Ullah, Caoilfhionn Lane, Teodora Buda, Brett Drury, Marc Mellotte, Haytham Assem, Michael Madden, Classification of cybercrime indicators in open social data, in: Proceedings of the 7th International Conference on Information Management and Big Data, Springer, 2020.

[7] Ihsan Ullah, Caoilfhionn Lane, Brett Drury, Marc Mellotte, Michael Madden, Open social data crime analytics, in: Proceedings of the International Workshop on Artificial Intelligence in Security, at IJCAI, Melbourne, Australia, 2017.

[8] Zhongqing Wang, Yue Zhang, DDoS event forecasting using Twitter data, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, AAAI Press, 2017, pp. 4151–4157.

[9] Yilun Wang, Michal Kosinski, Deep neural networks are more accurate than humans at detecting sexual orientation from facial images, J. Personal. Soc. Psychol. 114 (2) (2018) 246.

[10] Hal Berghel, Malice domestic: The Cambridge analytica dystopia, Computer 51 (5) (2018) 84–89.

[11] Felipe González, Yihan Yu, Andrea Figueroa, Claudia López, Cecilia Aragon, Global reactions to the cambridge analytica scandal: A cross-language social media study, in: Companion Proceedings of the 2019 World Wide Web Conference, 2019, pp. 799–806.

[12] Chinmayi Arun, Making choices: Social media platforms and freedom of expression norms, in: Regardless of Frontiers, 2018.

[13] Ogerta Elezaj, Sule Yildirim Yayilgan, Javed Ahmed, Edlira Kalemi, Brumle Brichfeld, Claudia Haubold, Crime intelligence from social media using CISMO, in: International Congress on Information and Communication Technology, Springer, 2020, pp. 441–460.

[14] Walid Salem, Chapter two NSAS and the possibility for transnational politics, in: Non-State Actors in Conflicts: Conspiracies, Myths, and Practices, Cambridge Scholars Publishing, 2018, p. 10.

[15] Crown Prosecution Service, Inchoate Offences, website, 2020, Accessed on October 17 2020. https://www.cps.gov.uk/legal-guidance/inchoate-offences.

[16] Jamila Boughelaf, Mobile Phones, Social Media and the Arab Spring, Technical report, Credemus Associates, 2011.

[17] Taylor Dewey, Juliane Kaden, Miriam Marks, Shun Matsushima, Beijing Zhu, The impact of social media on social unrest in the Arab Spring, Int. Policy Program 5 (8) (2012).

[18] Li Xiguang, Wang Jing, Web-based public diplomacy: The role of social media in the Iranian and Xinjiang riots, J. Int. Commun. 16 (1) (2010) 7–22.

[19] Daniel Briggs, Stephanie Alice Baker, From the criminal crowd to the "mediated crowd": the impact of social media on the 2011 English riots, in: Safer Communities, Emerald Group Publishing Limited, 2012.

[20] Caroline Rizza, Ângela Guimarães Pereira, Paula Curvelo, "Do-it-yourself justice": considerations of social media use in a crisis situation: the case of the 2011 vancouver riots, Int. J. Inf. Syst. Crisis Response Manage. (IJISCRAM) 6 (4) (2014) 42–59.

[21] Legislation.gov.uk, Computer Misuse Act 1990, website, 1990, Accessed 27/1/2022. https://www.legislation.gov.uk/ukpga/1990/18/contents.

[22] John Seymour, Philip Tully, Weaponizing data science for social engineering: Automated E2E spear phishing on Twitter, Black Hat USA 37 (2016).

[23] Michael Bossetta, The weaponization of social media: Spear phishing and cyberattacks on democracy, J. Int. Aff. 71 (1.5) (2018) 97–106.

[24] Daniel Gibert, Carles Mateu, Jordi Planes, The rise of machine learning for detection and classification of malware: Research developments, trends and challenges, J. Netw. Comput. Appl. 153 (2020) 102526.

[25] Weijie Han, Jingfeng Xue, Yong Wang, Zhenyan Liu, Zixiao Kong, MalInsight: A systematic profiling based malware detection framework, J. Netw. Comput. Appl. 125 (2019) 236–250.

[26] Bimal Parmar, Protecting against spear-phishing, Comput. Fraud Secur. 2012 (1) (2012) 8–11.

[27] Safwan Alam, Khalil El-Khatib, Phishing susceptibility detection through social media analytics, in: Proceedings of the 9th International Conference on Security of Information and Networks, 2016, pp. 61–64.

[28] Prateek Dewan, Anand Kashyap, Ponnurangam Kumaraguru, Analyzing social and stylometric features to identify spear phishing emails, in: 2014 Apwg Symposium on Electronic Crime Research (Ecrime), IEEE, 2014, pp. 1–13.

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,

[21] https://archive.fo/JdoOp.

Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[30] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei, Language models are few-shot learners, in: Proceedings of Advances in Neural Information Processing Systems, 2020.

[31] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, Quoc V. Le, XLNet: Generalized autoregressive pretraining for language understanding, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 5753–5763.

[32] David D. McDonald, Natural language generation., in: Handbook of Natural Language Processing, Vol. 2, 2010, pp. 121–144.

[33] Statute Law Database, Fraud Act 2006, website, 2006, Accessed 27/1/2022. https://www.legislation.gov.uk/ukpga/2006/35/contents.

[34] Mark Leiser, AstroTurfing,'CyberTurfing'and other online persuasion campaigns, Eur. J. Law Technol. 7 (1) (2016) 1–27.

[35] Marko Kovic, Adrian Rauchfleisch, Marc Sele, Christian Caspar, Digital astroturfing in politics: Definition, typology, and countermeasures, Stud. Commun. Sci. (2018).

[36] Jerry Zhang, Darrell Carpenter, Myung Ko, Online astroturfing: A theoretical perspective, in: AMCIS 2013 Proceedings, 2013.

[37] Kyumin Lee, Prithivi Tamilarasan, James Caverlee, Crowdturfers, campaigns, and social media: Tracking and revealing crowdsourced manipulation of social media, in: Seventh International AAAI Conference on Weblogs and Social Media, 2013.

[38] Kyumin Lee, Steve Webb, Hancheng Ge, Characterizing and automatically detecting crowdturfing in fiverr and Twitter, Soc. Netw. Anal. Min. 5 (1) (2015) 2.

[39] Jian Peng, Raymond Kim-Kwang Choo, Helen Ashman, Astroturfing detection in social media: Using binary n-gram analysis for authorship attribution, in: IEEE Trustcom/BigDataSE/ISPA, IEEE, 2016, pp. 121–128.

[40] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, Filippo Menczer, Detecting and tracking the spread of astroturf memes in microblog streams, 2010, CoRR, abs/1011.3768.

[41] Jonghyuk Song, Sangho Lee, Jong Kim, Crowdtarget: Target-based detection of crowdturfing in online social networks, in: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 2015, pp. 793–804.

[42] Yuanshun Yao, Bimal Viswanath, Jenna Cryan, Haitao Zheng, Ben Y. Zhao, Automated crowdturfing attacks and defenses in online review systems, in: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 2017, pp. 1143–1158.

[43] Albert Gatt, Emiel Krahmer, Survey of the state of the art in natural language generation: Core tasks, applications and evaluation, J. Artificial Intelligence Res. 61 (2018) 65–170.

[44] Victoria Baines, Fighting the industrialization of cybercrime, UN Chron. 50 (2) (2013) 10–12.

[45] Shuang Hao, Kevin Borgolte, Nick Nikiforakis, Gianluca Stringhini, Manuel Egele, Michael Eubanks, Brian Krebs, Giovanni Vigna, Drops for stuff: An analysis of reshipping mule scams, in: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 2015, pp. 1081–1092.

[46] Michael C. Galdo, Monica E. Tait, Lisa E. Feldman, Money mules: Stopping older adults and others from participating in international crime schemes, US Atty. Bull. 66 (2018) 95.

[47] Lisa Vaas, Get-rich-quick social media scams are turning teens into money mules, 2020, https://bit.ly/3aOIiJK. Online; accessed 11 May 2020.

[48] Brenda C. Arevalo, Money Mules: Facilitators of Financial Crime, Technical report, Utrecht University, 2015.

[49] Syed Mahbub, Eric Pardede, Using contextual features for online recruitment fraud detection, in: Proceedings of Designing Digitalization, 2018.

[50] Sangeeta Lal, Rishabh Jiaswal, Neetu Sardana, Ayushi Verma, Amanpreet Kaur, Rahul Mourya, ORFDetector: Ensemble learning based online recruitment fraud detection, in: Twelfth International Conference on Contemporary Computing (IC3), IEEE, 2019, pp. 1–5.

[51] Sokratis Vidros, Constantinos Kolias, Georgios Kambourakis, Online recruitment services: Another playground for fraudsters, Comput. Fraud Secur. 2016 (3) (2016) 8–13.

[52] Legislation.gov.uk, Terrorism Act 2000, website, 2000, Accessed 27/1/2022. https://archive.fo/hGoXl.

[53] Tinka Veldhuis, Jørgen Staun, Islamist Radicalisation: A Root Cause Model, Netherlands Institute of International Relations Clingendael The Hague, 2009.

[54] Ines Von Behr, Anaïs Reding, Charlie Edwards, Luke Gribbon, Radicalisation in the Digital Era: The Use of the Internet in 15 Cases of Terrorism and Extremism, Technical report, Rand Corporation, 2013.

[55] Gabriel Weimann, The emerging role of social media in the recruitment of foreign fighters, in: Foreign Fighters under International Law and beyond, Springer, 2016, pp. 77–95.

[56] Raúl Lara-Cabrera, Antonio Gonzalez-Pardo, David Camacho, Statistical analysis of risk assessment factors and metrics to evaluate radicalisation in Twitter, Future Gener. Comput. Syst. 93 (2019) 971–978.

[57] Swati Agarwal, Ashish Sureka, Using knn and svm based one-class classifier for detecting online radicalization on twitter, in: International Conference on Distributed Computing and Internet Technology, Springer, 2015, pp. 431–442.

[58] Oscar Araque, Carlos A. Iglesias, An approach for radicalization detection based on emotion signals and semantic similarity, IEEE Access 8 (2020) 17877–17891.

[59] Adam Bermingham, Maura Conway, Lisa McInerney, Neil O'Hare, Alan F. Smeaton, Combining social network analysis and sentiment analysis to explore the potential for online radicalisation, in: International Conference on Advances in Social Network Analysis and Mining, IEEE, 2009, pp. 231–236.

[60] Roger Bradbury, Terry Bossomaier, David Kernot, Predicting the emergence of self-radicalisation through social media: a complex systems approach, in: Terrorists' Use of the Internet: Assessment and Response, IOS Press, Amsterdam, 2017, pp. 379–389.

[61] Miriam Fernandez, Harith Alani, Contextual semantics for radicalisation detection on Twitter, in: Proceedings of Semantic Web for Social Good, CEUR, 2018.

[62] Emilio Ferrara, Wen-Qiang Wang, Onur Varol, Alessandro Flammini, Aram Galstyan, Predicting online extremism, content adopters, and interaction reciprocity, in: International Conference on Social Informatics, Springer, 2016, pp. 22–39.

[63] Matthew Rowe, Hassan Saif, Mining pro-ISIS radicalisation signals from social media users, in: Tenth International AAAI Conference on Web and Social Media, 2016.

[64] Hassan Saif, Thomas Dickinson, Leon Kastler, Miriam Fernandez, Harith Alani, A semantic graph-based approach for radicalisation detection on social media, in: European Semantic Web Conference, Springer, 2017, pp. 571–587.

[65] Ashish Sureka, Ponnurangam Kumaraguru, Atul Goyal, Sidharth Chhabra, Mining youtube to discover extremist videos, users and hidden communities, in: Asia Information Retrieval Symposium, Springer, 2010, pp. 13–24.

[66] Pooja Wadhwa, M.P.S. Bhatia, Measuring radicalization in online social networks using Markov Chains, J. Appl. Secur. Res. 10 (1) (2015) 23–47.

[67] Sarah Jane Delany, Pádraig Cunningham, Alexey Tsymbal, Lorcan Coyle, A case-based technique for tracking concept drift in spam filtering, in: International Conference on Innovative Techniques and Applications of Artificial Intelligence, Springer, 2004, pp. 3–16.

[68] Gabrielle Blanquart, David M. Cook, Twitter influence and cumulative perceptions of extremist support: A case study of geert wilders, in: Australian Counter Terrorism Conference, SRI Security Research Institute, Edith Cowan University, Perth, Western …, 2013.

[69] Miriam Fernandez, Antonio Gonzalez-Pardo, Harith Alani, Radicalisation influence in social media, Semant. Web J. (2019) In–Press.

[70] Queen's Printer of Acts of Parliament, Defamation Act 2013, website, 2013, Accessed 27/1/2022. https://archive.fo/bp21o.

[71] BAILII, England and Wales High Court (Queen's Bench Division) decisions - Soriano v Forensic News, website, 2021, Accessed 27/1/2022.

[72] Xinyi Zhou, Reza Zafarani, Fake news: A survey of research, detection methods, and opportunities, 2018, arXiv preprint arXiv:1812.00315.

[73] Julie Posetti, Alice Matthews, A Short Guide to the History of 'Fake News' and Disinformation, International Center for Journalists, 2018, pp. 01–19.

[74] Xinjiang Lu, Zhiwen Yu, Bin Guo, Xingshe Zhou, Predicting the content dissemination trends by repost behavior modeling in mobile social networks, J. Netw. Comput. Appl. 42 (2014) 197–207.

[75] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Alessandro Flammini, Filippo Menczer, The spread of fake news by social bots, 96 (2017) 104. arXiv preprint arXiv:1707.07592.

[76] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, David Lazer, Fake news on Twitter during the 2016 US presidential election, Science 363 (6425) (2019) 374–378.

[77] Meeyoung Cha, Wei Gao, Cheng-Te Li, Detecting fake news in social media: an Asia-Pacific perspective, Commun. ACM 63 (4) (2020) 68–71.

[78] Chuan Guo, Juan Cao, Xueyao Zhang, Kai Shu, Huan Liu, DEAN: Learning dual emotion for fake news detection on social media, 2019, arXiv preprint arXiv:1903.01728.

[79] Shengyi Jiang, Xiaoting Chen, Liming Zhang, Sutong Chen, Haonan Liu, User-characteristic enhanced model for fake news detection in social media, in: CCF International Conference on Natural Language Processing and Chinese Computing, Springer, 2019, pp. 634–646.

[80] Rahul Mishra, Vinay Setty, SADHAN: Hierarchical attention networks to learn latent aspect embeddings for fake news detection, in: Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval, 2019, pp. 197–204.

[81] Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, Michael M. Bronstein, Fake news detection on social media using geometric deep learning, 2019, arXiv preprint arXiv:1902.06673.

[82] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, Huan Liu, Fake news detection on social media: A data mining perspective, ACM SIGKDD Explor. Newsl. 19 (1) (2017) 22–36.

[83] Kai Shu, Suhang Wang, Huan Liu, Beyond news contents: The role of social context for fake news detection, in: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, 2019, pp. 312–320.

[84] Liang Wu, Huan Liu, Tracing fake-news footprints: Characterizing social media messages by how they propagate, in: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, 2018, pp. 637–645.

[85] Gerald Ki Wei Huang, Jun Choi Lee, Hyperpartisan news and articles detection using BERT and ELMo, in: International Conference on Computer and Drone Applications (IConDA), IEEE, 2019, pp. 29–32.

[86] Yang Yang, Lei Zheng, Jiawei Zhang, Qingcai Cui, Zhoujun Li, Philip S. Yu, TI-CNN: Convolutional neural networks for fake news detection, 2018, arXiv preprint arXiv:1806.00749.

[87] Crown Prosecution Service, Social Media - Guidelines on Prosecuting Cases Involving Communications Sent via Social Media, website, 2018, Accessed 27/1/2022. https://archive.fo/uMkZc.

[88] Cristan Williams, Radical inclusion: Recounting the trans inclusive history of radical feminism, Transgender Stud. Q. 3 (1–2) (2016) 254–258.

[89] Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, Ingmar Weber, Analyzing the targets of hate in online social media, in: Tenth International AAAI Conference on Web and Social Media, 2016.

[90] Thais G. Almeida, Fabíola G. Nakamura, Eduardo F. Nakamura, Uma abordagem para identificar e monitorar haters em redes sociais online, in: Anais do XXIII Simpósio Brasileiro de Sistemas Multimídia e Web, 2017.

[91] Pete Burnap, Matthew L. Williams, Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making, Policy Internet 7 (2) (2015) 223–242.

[92] Pete Burnap, Matthew L. Williams, Us and them: identifying cyber hate on Twitter across multiple protected characteristics, EPJ Data Sci. 5 (1) (2016).

[93] Michael Chau, Jennifer Xu, Mining communities and their relationships in blogs: A study of online hate groups, Int. J. Hum.-Comput. Stud. 65 (1) (2007) 57–70.

[94] Hsinchun Chen, Sven Thoms, Tianjun Fu, Cyber extremism in Web 2.0: An exploratory study of international Jihadist groups, in: Intelligence and Security Informatics, 2008. ISI 2008. IEEE International Conference on, IEEE, 2008, pp. 98–103.

[95] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, Narayan Bhamidipati, Hate speech detection with comment embeddings, in: Proceedings of the 24th International Conference on World Wide Web, ACM, 2015, pp. 29–30.

[96] Michele P. Hamm, Amanda S. Newton, Annabritt Chisholm, Jocelyn Shulhan, Andrea Milne, Purnima Sundar, Heather Ennis, Shannon D. Scott, Lisa Hartling, Prevalence and effect of cyberbullying on children and young people: A scoping review of social media studies, JAMA Pediatr. 169 (8) (2015) 770–777.

[97] Sweta Agrawal, Amit Awekar, Deep learning for detecting cyberbullying across multiple social media platforms, in: European Conference on Information Retrieval, Springer, 2018, pp. 141–153.

[98] Maral Dadvar, Kai Eckert, Cyberbullying detection in social networks using deep learning based models; a reproducibility study, 2018, arXiv preprint arXiv:1812.08046.

[99] Maral Dadvar, F.M.G. de Jong, Roeland Ordelman, Dolf Trieschnigg, Improved cyberbullying detection using gender information, in: Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012), University of Ghent, 2012.

[100] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, Franciska de Jong, Improving cyberbullying detection with user context, in: European Conference on Information Retrieval, Springer, 2013, pp. 693–696.

[101] Harsh Dani, Jundong Li, Huan Liu, Sentiment informed cyberbullying detection in social media, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2017, pp. 52–67.

[102] Karthik Dinakar, Roi Reichart, Henry Lieberman, Modeling the detection of textual cyberbullying, in: Fifth International AAAI Conference on Weblogs and Social Media, 2011.

[103] Rui Zhao, Anna Zhou, Kezhi Mao, Automatic detection of cyberbullying on social networks based on bullying features, in: Proceedings of the 17th International Conference on Distributed Computing and Networking, 2016, pp. 1–6.

[104] Mohammed Ali Al-garadi, Kasturi Dewi Varathan, Sri Devi Ravana, Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network, Comput. Hum. Behav. 63 (2016) 433–443.

[105] Chuong B. Do, Andrew Y. Ng, Transfer learning for text classification, in: Advances in Neural Information Processing Systems, 2006, pp. 299–306.

[106] Adam D.I. Kramer, Jamie E. Guillory, Jeffrey T. Hancock, Experimental evidence of massive-scale emotional contagion through social networks, Proc. Natl. Acad. Sci. 111 (24) (2014) 8788–8790.

[107] Shashank Srivastava, Igor Labutov, Tom Mitchell, Zero-shot learning of classifiers from natural language quantification, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 306–316.

[108] Alan Ramponi, Barbara Plank, Neural unsupervised domain adaptation in NLP—A survey, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 6838–6855.

[109] Jiaoyan Chen, Yuxia Geng, Zhuo Chen, Ian Horrocks, Jeff Z. Pan, Huajun Chen, Knowledge-aware zero-shot learning: Survey and perspective, in: Proceedings of the 2021 Conference of International Joint Conference on Artificial Intelligence, 2021.

[110] Sinno Jialin Pan, Qiang Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2009) 1345–1359.

[111] Emily Allaway, Malavika Srikanth, Kathleen Mckeown, Adversarial learning for zero-shot stance detection on social media, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 4756–4767.

[112] Karl Weiss, Taghi M. Khoshgoftaar, DingDing Wang, A survey of transfer learning, J. Big Data 3 (1) (2016) 1–40.

[113] Marzieh Mozafari, Reza Farahbakhsh, Noel Crespi, A BERT-based transfer learning approach for hate speech detection in online social media, in: International Conference on Complex Networks and their Applications, Springer, 2019, pp. 928–940.

[114] Huina Mao, Xin Shuai, Apu Kapadia, Loose tweets: an analysis of privacy leaks on twitter, in: Proceedings of the 10th Annual ACM Workshop on Privacy in the Electronic Society, ACM, 2011, pp. 1–12.

[115] Ieng-Fat Lam, Kuan-Ta Chen, Ling-Jyh Chen, Involuntary information leakage in social network services, in: International Workshop on Security, Springer, 2008, pp. 167–183.

[116] Nurul Nuha Abdul Molok, Atif Ahmad, Shanton Chang, et al., Information leakage through online social networking: Opening the doorway for advanced persistence threats, J. Aust. Inst. Prof. Intell. Off. 19 (2) (2011) 38.

[117] Katharina Krombholz, Heidelinde Hobel, Markus Huber, Edgar Weippl, Advanced social engineering attacks, J. Inf. Secur. Appl. 22 (2015) 113–122.

[118] M.F. Vilardo, Online impersonation in securities scams, IEEE Secur. Priv. 2 (3) (2004) 82–85.

[119] Nurul Nuha Abdul Molok, Atif Ahmad, Shanton Chang, Online social networking: a source of intelligence for advanced persistent threats, Int. J. Cyber Warf. Terror. (IJCWT) 2 (1) (2012) 1–13.

[120] Danesh Irani, Steve Webb, Kang Li, Calton Pu, Modeling unintended personal-information leakage from multiple online social networks, IEEE Internet Comput. 15 (3) (2011) 13–19.

[121] Terence Chen, Mohamed Ali Kaafar, Arik Friedman, Roksana Boreli, Is more always merrier?: a deep dive into online social footprints, in: Proceedings of the 2012 ACM Workshop on Workshop on Online Social Networks, ACM, 2012, pp. 67–72.

[122] Frank L. Greitzer, Ryan E. Hohimer, Modeling human behavior to anticipate insider attacks, J. Strateg. Secur. 4 (2) (2011) 25.

[123] Oana Goga, Daniele Perito, Howard Lei, Renata Teixeira, Robin Sommer, Large-Scale Correlation of Accounts across Social Networks, Tech. Rep. TR-13-002, University of California at Berkeley, Berkeley, California, 2013.

[124] Paridhi Jain, Ponnurangam Kumaraguru, Anupam Joshi, @ I seek'fb. me': Identifying users across multiple online social networks, in: Proceedings of the 22nd International Conference on World Wide Web, ACM, 2013, pp. 1259–1268.

[125] Yuanping Nie, Yan Jia, Shudong Li, Xiang Zhu, Aiping Li, Bin Zhou, Identifying users across social networks based on dynamic core interests, Neurocomputing 210 (2016) 107–115.

[126] Alexander Panchenko, Dmitry Babaev, Sergei Obiedkov, Large-scale parallel matching of social network profiles, in: International Conference on Analysis of Images, Social Networks and Texts, Springer, 2015, pp. 275–285.

[127] Katerina Zamani, Georgios Paliouras, Dimitrios Vogiatzis, Similarity-based user identification across social networks, in: International Workshop on Similarity-Based Pattern Recognition, Springer, 2015, pp. 171–185.

[128] Aunshul Rege, What's love got to do with it? Exploring online dating scams and identity fraud, Int. J. Cyber Criminol. 3 (2) (2009).

[129] M.J. Berry, S.L. Bainbridge, Manchester's Cyberstalked 18-30s: Factors affecting cyberstalking, Adv. Soc. Sci. Res. J. 4 (18) (2017).

[130] Muhammad Al-Qurishi, Mabrook Al-Rakhami, Atif Alamri, Majed Alrubaian, Sk. Md. Mizanur Rahman, M. Shamim Hossain, Sybil defense techniques in online social networks: a survey, IEEE Access 5 (2017) 1200–1219.

[131] Mark F. Vilardo, Online impersonation in securities scams, IEEE Secur. Priv. 2 (3) (2004) 82–85.

[132] Arijit De, Colin M. Bogart, Caitlin S. Collins, Detecting impersonation on a social network, 2015, US Patent 9, 224, 008.

[133] Brad S. Trinkle, Robert E. Crossler, Merrill Warkentin, I'm game, are you? Reducing real-world security threats by managing employee activity in online social networks, J. Inf. Syst. 28 (2) (2014) 307–327.

[134] Surette Ray, How social media is changing the way people commit crimes andpolice fight them, in: USApp–American Politics and Policy Blog, The London School of Economics and Political Science, 2016.

[135] Mohammad Al Boni, Matthew S. Gerber, Predicting crime with routine activity patterns inferred from social media, in: IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, 2016, pp. 001233–001238.

[136] Rasim M. Alguliyev, Ramiz M. Aliguliyev, Fargana J. Abdullayeva, Deep learning method for prediction of DDoS attacks on social media, Adv. Data Sci. Adapt. Anal. 11 (01n02) (2019) 1950002.

[137] Nasser Alsaedi, Pete Burnap, Omer Rana, Can we predict a riot? Disruptive event detection using Twitter, ACM Trans. Internet Technol. (TOIT) 17 (2) (2017) 1–26.

[138] Elhadj Benkhelifa, Elliott Rowe, Robert Kinmond, Oluwasegun A. Adedugbe, Thomas Welsh, Exploiting social networks for the prediction of social and civil unrest: A cloud based framework, in: International Conference on Future Internet of Things and Cloud, IEEE, 2014, pp. 565–572.

[139] Jose Cadena, Gizem Korkmaz, Chris J. Kuhlman, Achla Marathe, Naren Ramakrishnan, Anil Vullikanti, Forecasting social unrest using activity cascades, PLoS One 10 (6) (2015).

[140] Xinyu Chen, Youngwoon Cho, Suk Young Jang, Crime prediction using twitter sentiment and weather, in: Systems and Information Engineering Design Symposium, IEEE, 2015, pp. 63–68.

[141] Coral Featherstone, The relevance of social media as it applies in South Africa to crime prediction, in: IST-Africa Conference & Exhibition, IEEE, 2013, pp. 1–7.

[142] Coral Featherstone, Identifying vehicle descriptions in microblogging text with the aim of reducing or predicting crime, in: International Conference on Adaptive Science and Technology, IEEE, 2013, pp. 1–8.

[143] Matthew S. Gerber, Predicting crime using Twitter and kernel density estimation, Decis. Support Syst. 61 (2014) 115–125.

[144] Rob Procter, Farida Vis, Alex Voss, Reading the riots on Twitter: methodological innovation for the analysis of big data, Int. J. Soc. Res. Methodol. 16 (3) (2013) 197–214.

[145] Aldo Hernandez-Suarez, Gabriel Sanchez-Perez, Karina Toscano-Medina, Victor Martinez-Hernandez, Hector Perez-Meana, Jesus Olivares-Mercado, Victor Sanchez, Social sentiment sensor in Twitter for predicting cyber-attacks using l1 regularization, Sensors 18 (5) (2018) 1380.

[146] Mingjun Wang, Matthew S. Gerber, Using twitter for next-place prediction, with an application to crime prediction, in: IEEE Symposium Series on Computational Intelligence, IEEE, 2015, pp. 941–948.

[147] Xiaofeng Wang, Matthew S. Gerber, Donald E. Brown, Automatic crime prediction using events extracted from twitter posts, in: International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, Springer, 2012, pp. 231–238.

[148] Jeffery Ansah, Wei Kang, Lin Liu, Jixue Liu, Jiuyong Li, Information propagation trees for protest event prediction, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2018, pp. 777–789.

[149] Mohsen Bahrami, Yasin Findik, Burcin Bozkaya, Selim Balcisoy, Twitter reveals: Using Twitter analytics to predict public protests, 2018, arXiv preprint arXiv:1805.00358.

[150] Rostyslav Korolov, Di Lu, Jingjing Wang, Guangyu Zhou, Claire Bonial, Clare Voss, Lance Kaplan, William Wallace, Jiawei Han, Heng Ji, On predicting social unrest using social media, in: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2016, pp. 89–95.

[151] Shannon C. McGregor, Social media as public opinion: How journalists use social media to represent public opinion, Journalism 20 (8) (2019) 1070–1086.

[152] Georgios Paltoglou, Sentiment analysis in social media, in: Online Collective Action, Springer, 2014, pp. 3–17.

[153] Alan Mishler, Kevin Wonus, Wendy Chambers, Michael Bloodgood, Filtering tweets for social unrest, in: IEEE 11th International Conference on Semantic Computing (ICSC), IEEE, 2017, pp. 17–23.

[154] National Academies of Sciences Engineering, Proactive Policing: Effects on Crime and Communities, National Academies Press, 2018.

[155] Joseph Campbell Jr., Alyssa C. Mensch, Giselle Zeno, William M. Campbell, Richard P. Lippmann, David J. Weller-Fahy, Finding Malicious Cyber Discussions in Social Media, Technical report, MIT Lincoln Laboratory, Lexington United States, 2015.

[156] Ogerta Elezaj, Sule Yildirim Yayilgan, Edlira Kalemi, Criminal network community detection in social media forensics, in: International Conference on Intelligent Technologies and Applications, Springer, 2020, pp. 371–383.

[157] Zenun Kastrati, Ali Shariq Imran, Sule Yildirim-Yayilgan, Fisnik Dalipi, Analysis of online social networks posts to investigate suspects using SEMCON, in: International Conference on Social Computing and Social Media, Springer, 2015, pp. 148–157.

[158] Edlira Kalemi, Sule Yildirim-Yayilgan, Elton Domnori, Ogerta Elezaj, SMONT: an ontology for crime solving through social media, Int. J. Metadata Semant. Ontol. 12 (2–3) (2017) 71–81.

[159] Henrik Legind Larsen, José María Blanco, Raquel Pastor Pastor, Ronald R. Yager, Using Open Data to Detect Organized Crime Threats: Factors Driving Future Crime, Springer, 2017.

[160] Raymond Y.K. Lau, Yunqing Xia, Yunming Ye, A probabilistic generative model for mining cybercriminal networks from online social media, IEEE Comput. Intell. Mag. 9 (1) (2014) 31–43.

[161] Richard P. Lippmann, Joseph P. Campbell, David J. Weller-Fahy, Alyssa C. Mensch, William M. Campbell, Finding Malicious Cyber Discussions in Social Media, Technical report, MIT Lincoln Laboratory, Lexington United States, 2016.

[162] Soufiane Maguerra, Azedine Boulmakoul, Lamia Karim, Hassan Badir, Scalable solution for profiling potential cyber-criminals in Twitter, in: Proceedings of the Big Data & Applications 12th Edition of the Conference on Advances of Decisional Systems, Marrakech, Morocco, 2018, pp. 2–3.

[163] Christopher C. Yang, Tobun D. Ng, Terrorism and crime related weblog social network: Link, content analysis and information visualization, in: IEEE Intelligence and Security Informatics, IEEE, 2007, pp. 55–58.

[164] Jon Stone, New law requiring ID cards to open social media accounts debated in Italy, 2020, https://bit.ly/2VYAM9v. Online; accessed 19 April 2020.

[165] Supraja Gurajala, Joshua S. White, Brian Hudson, Brian R. Voter, Jeanna N. Matthews, Profile characteristics of fake Twitter accounts, Big Data Soc. 3 (2) (2016) 2053951716674236.

[166] Osnat Mokryn, Hagit Ben-Shoshan, Domain-based latent personal analysis and its use for impersonation detection in social media, User Model. User-Adapt. Interact. 31 (4) (2021) 785–828.

[167] Oana Goga, Giridhari Venkatadri, Krishna P. Gummadi, The doppelgänger bot attack: Exploring identity impersonation in online social networks, in: Proceedings of the 2015 Internet Measurement Conference, 2015, pp. 141–153.

[168] Akshay J. Sarode, Arun Mishra, Audit and analysis of impostors: An experimental approach to detect fake profile in online social network, in: Proceedings of the Sixth International Conference on Computer and Communication Technology 2015, 2015, pp. 1–8.

[169] Estée Van Der Walt, Jan Eloff, Using machine learning to detect fake identities: bots vs humans, IEEE Access 6 (2018) 6540–6549.

[170] Estee Van der Walt, Jan H.P. Eloff, Jacomine Grobler, Cyber-security: Identity deception detection on social media platforms, Comput. Secur. 78 (2018) 76–89.

[171] Koosha Zarei, Reza Farahbakhsh, Noel Crespi, Deep dive on politician impersonating accounts in social media, in: Proceedings of the 24th Symposium on Computers and Communications, 2019, pp. 1–6.

[172] Koosha Zarei, Reza Farahbakhsh, Noël Crespi, Typification of impersonated accounts on instagram, in: IEEE 38th International Performance Computing and Communications Conference (IPCCC), IEEE, 2019, pp. 1–6.

[173] Haipeng Chen, Jing Liu, Rui Liu, Noseong Park, V.S. Subrahmanian, VASE: A Twitter-based vulnerability analysis and score engine, in: Proceedings of the International Conference on Data Mining, 2019.

[174] Linqing Liu, Yao Lu, Ye Luo, Renxian Zhang, Laurent Itti, Jianwei Lu, Detecting “Smart” spammers on social network: A topic model approach, in: Proceedings of the NAACL Student Research Workshop, 2016, pp. 45–50.

[175] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, Huan Liu, Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media, J. Big Data 8 (3) (2020).

[176] Thomas Davidson, Dana Warmsley, Michael Macy, Ingmar Weber, Automated hate speech detection and the problem of offensive language, in: Proceedings of the 11th International AAAI Conference on Web and Social Media, in: ICWSM ’17, 2017, pp. 512–515.

[177] Polychronis Charitidis, Stavros Doropoulos, Stavros Vologiannidis, Ioannis Papastergiou, Sophia Karakeva, Hate speech and personal attack dataset in German social media, 2019.

[178] Polychronis Charitidis, Stavros Doropoulos, Stavros Vologiannidis, Ioannis Papastergiou, Sophia Karakeva, Hate Speech and Personal Attack Dataset in Spanish Social Media, website, 2019.

[179] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, Ritesh Kumar, Predicting the type and target of offensive posts in social media, in: Proceedings of NAACL, 2019.

[180] Sokratis Vidros, Constantinos Kolias, Georgios Kambourakis, Leman Akoglu, Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset, Future Internet 9 (1) (2017) 6.

[181] Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kadras, Sylvain Gugger, Jeremy Howard, MultiFiT: Efficient multi-lingual language model fine-tuning, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 5702–5707.