

# The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans

Savvas Zannettou<sup>\*</sup>, Michael Sirivianos<sup>\*</sup>, Jeremy Blackburn<sup>‡</sup>, Nicolas Kourtellis<sup>†</sup>

<sup>\*</sup>Cyprus University of Technology, <sup>‡</sup>University of Alabama at Birmingham, <sup>†</sup>Telefonica Research  
sa.zannettou@edu.cut.ac.cy, michael.sirivianos@cut.ac.cy, blackburn@uab.edu, nicolas.kourtellis@telefonica.com

## Abstract

A new era of Information Warfare has arrived. Various actors, including state-sponsored ones, are weaponizing information on Online Social Networks to run false information campaigns with targeted manipulation of public opinion on specific topics. These false information campaigns can have dire consequences to the public: mutating their opinions and actions, especially with respect to critical world events like major elections. Evidently, the problem of false information on the Web is a crucial one, and needs increased public awareness, as well as immediate attention from law enforcement agencies, public institutions, and in particular, the research community.

In this paper, we make a step in this direction by providing a typology of the Web's false information ecosystem, comprising various types of false information, actors, and their motives. We report a comprehensive overview of existing research on the false information ecosystem by identifying several lines of work: 1) how the public perceives false information; 2) understanding the propagation of false information; 3) detecting and containing false information on the Web; and 4) false information on the political stage. In this work, we pay particular attention to political false information as: 1) it can have dire consequences to the community (e.g., when election results are mutated) and 2) previous work show that this type of false information propagates faster and further when compared to other types of false information. Finally, for each of these lines of work, we report several future research directions that can help us better understand and mitigate the emerging problem of false information dissemination on the Web.

## 1 Introduction

Online Social Networks (OSNs) play an important role in the way people communicate and consume information. This is mainly because OSNs provide an ideal environment for communication and information acquisition, as users have access to a staggering amount of posts and articles that can share with others in real-time. Unfortunately, OSNs have also become the mechanism for massive campaigns to diffuse false information [15, 6]. In particular, recent reporting has highlighted how OSNs are exploited by powerful actors, potentially even state level, in order to manipulate individuals via targeted disinformation campaigns [12, 13].

The extensive dissemination of false information in OSNs can pose a major problem, affecting society in extremely worrisome ways. For example, false information can hurt the image of a candidate, potentially altering the outcome of an election. During crisis situations (e.g., terrorist attacks, earthquakes, etc.), false information can cause wide spread panic and general chaos (e.g., [8]). False information diffusion in OSNs is achieved via diverse types of users, which typically have various motives. A diverse set of users are involved in the diffusion of false information, some unwittingly, and some with particular motives. For example, terrorist organizations exploit OSNs to deliberately diffuse false information for propaganda purposes [28]. Malicious users might utilize sophisticated automation tools (i.e., bots) or fake accounts that target specific benign users with the goal of influencing ideology. No matter the motivation, however, the effects false information has on society clearly indicate the need for better understanding, measurement, and mitigation of false information in OSNs.

In this work, we provide a typology of the false information ecosystem that sheds light on the following questions: 1) What are the various types and instances of false information on the Web? 2) Who are the different actors that diffuse false information on the Web? and 3) What are the motives behind the spread of false information? Our typology is built after an extensive study of the existing literature, where we identify the following lines of work regarding the false information ecosystem:

- **User perception of false information.** This refers to how users perceive and interact with false information that is disseminated in OSNs. For example, can users distinguish real stories from fake? If yes, what cues do they use to make this distinction?
- **Propagation dynamics of false information in OSNs.** Understanding the underlying propagation dynamics provides useful insights regarding the problem that can assist in detecting and mitigating false information. In this line of work, we will review studies that focus on the propagation aspect of false information without expanding to detection and containment techniques.

- **Detection and containment of false information.** Detecting and containing the propagation of false information is a desired outcome. However, no robust platform, service or system is in-place that can effectively and efficiently mitigate the problem in a timely manner. In this line of work, we will provide an overview of the community’s efforts for detecting and containing the spread of false information on the Web.
- **False information in politics.** This line of work refers to work that focus on politics-related false information. Note, that this line of work overlaps with all of the other lines of work; however, we elect to devote a separate line of work for this survey for several reasons. First, anecdotal evidence and existing literature suggest that false information is often disseminated for politics-related reasons. Second, it can affect the course of history, especially in election periods. For instance, during the 2016 US Election, as well as the 2017 French Elections, there were examples where trolls from 4chan tried to change the election outcome by disseminating false information about the candidate they opposed [9, 24]. On top of this, extensive anecdotal evidence suggests the involvement of state-sponsored troll accounts that actively try to mislead and mutate the opinions of users on OSNs [172, 201]. Also, previous work has shown that politics-related false information have a more intensive effect on social media when compared to other types of false information [181]. Last, some of the previous work is specific to political context and is outside of the scope of the other three lines of work.

For each of these lines of work, we provide an overview of the most relevant research papers, as well as possible future research directions that will address existing gaps and will better help the community to alleviate the emerging problem of false information on the Web. Note that the main focus of our literature review is towards computational approaches in understanding and tackling the problem of false information on the Web; however, we also report some previous work that sheds light on some other interesting features like interface design, socio-technical aspects, as well as systems that can help mitigate the problem.

**Contributions.** In summary, with the present study we make the following contributions: First, we provide a general overview of the false information ecosystem by proposing a typology. This will help other researchers that want to work on this field and have no previous experience with the topic. Second, we provide a comprehensive overview of the related work that fall in one of the identified lines of work. This overview can act as an index for other researchers, which will be able to quickly identify previous work on a thematic, what methodology was used, and what Web communities previous work considered. Furthermore, it can assist other researchers to quickly identify gaps in the literature. Note that we acknowledge that the information ecosystem is huge and complex, hence we elect to have a broad scope in our survey, while paying particular attention to the political aspect of false information. This is mainly because it can have alarming consequences on the community and because previous work shows that the effects of false information are more intense in political contexts when compared to other types [181]. Finally, for each identified line of work, we provide some future directions that the community can follow to extend the state-of-art on the Web’s information ecosystem.

**Paper Organization.** The remainder of this paper is structured as follows: In Section 2 we describe a typology for the false information ecosystem. Section 3 reports existing work on studying how user perceive and interact with false information on the Web, while Section 4 describes the studies done on the propagation of information. Studies about detecting and containing false information on the Web are presented in Section 5. Section 6 is devoted on false information in the politics stage whereas Section 7 reports other relevant work that does not fit in any of the other lines of work. Finally, we conclude in Section 8.

## 2 False Information Ecosystem Typology

In this section we present our typology, which we believe it will provide a succinct roadmap for future work. The typology is based on [7] and extended to build upon the existing literature. Specifically, we describe the various types of false information that can be found in OSNs (Section 2.1), the various types of actors that contribute in the distribution of false information (Section 2.2), as well as their motives (Section 2.3). Note that our typology is different from concurrent work by Kumar and Shah [108] as we provide a fine-grained distinction for the types of false information, the actors, and their motives. Also, note that we make a best effort to cover as many aspects of the false information as per our knowledge; however, the typology should not be treated as an exhaustive representation of the false information ecosystem.

### 2.1 Types of False Information

False information on the Web can be found in various forms, hence we propose the categorization of false information into eight types as listed below:

- **Fabricated (F) [154].** Completely fictional stories disconnected entirely from real facts. This type is not new and it exists since the birth of journalism. Some popular examples include fabricated stories about politicians and aliens [84] (e.g., the story that Hillary Clinton adopted an alien baby).
- **Propaganda (P) [101].** This is a special instance of the fabricated stories that aim to harm the interests of a particular party and usually has a political context. This kind of false news is not new, as it was widely used during World War II

and the Cold War. Propaganda stories are profoundly utilized in political contexts to mislead people with the overarching goal of inflicting damage to a particular political party or nation-state. Due to this, propaganda is a consequential type of false information as it can change the course of human history (e.g., by changing the outcome of an election). Some recent examples of propaganda include stories about the Syria air strikes in 2018 or about specific movements like the BlackLivesMatter (see [124] for more examples).

- **Conspiracy Theories (CT) [69]**. Refer to stories that try to explain a situation or an event by invoking a conspiracy without proof. Usually, such stories are about illegal acts that are carried out by governments or powerful individuals. They also typically present unsourced information as fact or dispense entirely with an “evidence” based approach, relying on leaps of faith instead. Popular recent examples of conspiracy theories include the Pizzagate theory (i.e., Clinton’s campaign running a pedophile ring) [188] and conspiracies around the murder of Seth Rich [187] (e.g., Seth Rich was involved in the DNC email leaks).
- **Hoaxes (H) [109]**. News stories that contain facts that are either false or inaccurate and are presented as legitimate facts. This category is also known in the research community either as half-truth [3] or factoid [14] stories. Popular examples of hoaxes are stories that report the false death of celebrities (e.g., the Adam Sadler death hoax [164]).
- **Biased or one-sided (B)**. Refers to stories that are extremely one-sided or biased. In the political context, this type is known as Hyperpartisan news [144] and are stories that are extremely biased towards a person/party/situation/event. Some examples include the wide spread diffusion of false information to the alt-right community from small fringe Web communities like 4chan’s /pol/ board [86] and Gab, an alt-right echo chamber [199].
- **Rumors (R) [140]**. Refers to stories whose truthfulness is ambiguous or never confirmed. This kind of false information is widely propagated on OSNs, hence several studies have analyzed this type of false information. Some examples of rumors include stories around the 2013 Boston Marathon Bombings like the story that the suspects became citizens on 9/11 or that a Sandy Hook child was killed during the incident [163].
- **Clickbait (CL) [56]**. Refers to the deliberate use of misleading headlines and thumbnails of content on the Web. This type is not new as it appeared years before, during the “newspaper era,” a phenomenon known as yellow journalism [50]. However, with the proliferation of OSNs, this problem is rapidly growing, as many users add misleading descriptors to their content with the goal of increasing their traffic for profit or popularity [143]. This is one of the least severe types of false information because if a user reads/views the whole content then he can distinguish if the headline and/or the thumbnail was misleading.
- **Satire News (S) [48]**. Stories that contain a lot of irony and humor. This kind of news is getting considerable attention on the Web in the past few years. Some popular examples of sites that post satire news are TheOnion [22] and SatireWire [16]. Usually, these sites disclose their satyric nature in one of their pages (i.e., About page). However, as their articles are usually disseminated via social networks, this fact is obfuscated, overlooked, or ignored by users who often take them at face value with no additional verification.

It is extremely important to highlight that there is an overlap in the aforementioned types of false information, thus it is possible to observe false information that may fall within multiple categories. Here, we list two indicative examples to better understand possible overlaps: 1) a rumor may also use clickbait techniques to increase the audience that will read the story; and 2) propaganda stories, which are a special instance of a fabricated story, may also be biased towards a particular party. These examples highlight that the false information ecosystem is extremely complex and the various types of false information need to be considered to mitigate the problem.

## 2.2 False Information Actors

In this section, we describe the different types of actors that constitute the false information propagation ecosystem. We identified a handful of different actors that we describe below.

- **Bots [43]**. In the context of false information, bots are programs that are part of a bot network (Botnet) and are responsible for controlling the online activity of several fake accounts with the aim of disseminating false information. Botnets are usually tied to a large number of fake accounts that are used to propagate false information in the wild. A Botnet is usually employed for profit by 3rd party organizations to diffuse false information for various motives (see Section 2.3 for more information on their possible motives). Note that various types of bots exist, which have varying capabilities; for instance, some bots only repost content, promote content (e.g., via vote manipulation on Reddit or similar platforms), and others post “original” content. However, this distinction is outside of the scope of this work, which provides a general overview of the information ecosystem on the Web.

- **Criminal/Terrorist Organizations [28].** Criminal gangs and terrorist organizations are exploiting OSNs as the means to diffuse false information to achieve their goals. A recent example is the ISIS terrorist organization that diffuses false information in OSNs for propaganda purposes [28]. Specifically, they widely diffuse ideologically passionate messages for recruitment purposes. This creates an extremely dangerous situation for the community as there are several examples of individuals from European countries recruited by ISIS that ended-up perpetrating terrorist acts.
- **Activist or Political Organizations.** Various organizations share false information in order to either promote their organization, demote other rival organizations, or for pushing a specific narrative to the public. A recent example include the National Rifle Association, a non-profit organization that advocates gun rights, which disseminated false information to manipulate people about guns [173]. Other examples include political parties that share false information, especially near major elections [29].
- **Governments [175].** Historically, governments were involved in the dissemination of false information for various reasons. More recently, with the proliferation of the Internet, governments utilize the social media to manipulate public opinion on specific topics. Furthermore, there are reports that foreign governments share false information on other countries in order to manipulate public opinion on specific topics that regard the particular country. Some examples, include the alleged involvement of the Russian government in the 2016 US elections [159] and Brexit referendum [152].
- **Hidden Paid Posters [54] and State-sponsored Trolls [201].** They are a special group of users that are paid in order to disseminate false information on a particular content or targeting a specific demographic. Usually, they are employed for pushing an agenda; e.g., to influence people to adopt certain social or business trends. Similar to bots, these actors disseminate false information for profit. However, this type is substantially harder to distinguish than bots because they exhibit characteristics similar to regular users.
- **Journalists [114].** Individuals that are the primary entities responsible for disseminating information both to the online and to the offline world. However, in many cases, journalists are found in the center of controversy as they post false information for various reasons. For example, they might change some stories so that they are more appealing, in order to increase the popularity of their platform, site, or newspaper.
- **Useful Idiots [23].** The term originates from the early 1950s in the USA as a reference to a particular political party's members that were manipulated by Russia in order to weaken the USA. Useful idiots are users that share false information mainly because they are manipulated by the leaders of some organization or because they are naive. Usually, useful idiots are normal users that are not fully aware of the goals of the organization, hence it is extremely difficult to identify them. Like hidden paid posters, useful idiots are hard to distinguish and there is no study that focuses on this task.
- **“True Believers” and Conspiracy Theorists.** Refer to individuals that share false information because they actually believe that they are sharing the truth and that other people need to know about it. For instance, a popular example is Alex Jones, which is a popular conspiracy theorist that shared false information about the Sandy Hook shooting [189].
- **Individuals that benefit from false information.** Refer to various individuals that will have a personal gain by disseminating false information. This is a very broad category ranging from common persons like an owner of a cafeteria to popular individuals like political persons.
- **Trolls [126].** The term troll is used in great extent by the Web community and refers to users that aim to do things to annoy or disrupt other users, usually for their own personal amusement. An example of their arsenal is posting provocative or off-topic messages in order to disrupt the normal operation or flow of discussion of a website and its users. In the context of false information propagation, we define trolls as users that post controversial information in order to provoke other users or inflict emotional pressure. Traditionally, these actors use fringe communities like Reddit and 4chan to orchestrate organized operations for disseminating false information to mainstream communities like Twitter, Facebook, and YouTube [200, 86].

Similarly to the types of false information, overlap may exist in actors too. Some examples include: 1) Bots can be exploited by criminal organizations or political persons to disseminate false information [11]; and 2) Hidden paid posters and state-sponsored trolls can be exploited by political persons or organizations to push false information for a particular agenda [12].

### 2.3 Motives behind false information propagation

False information actors and types have different motives behind them. Below we describe the categorization of motives that we distinguish:

- **Malicious Intent.** Refers to a wide spectrum of intents that drive actors that want to hurt others in various ways. Some examples include inflicting damage to the public image of a specific person, organization, or entity.

Platform	OSN data analysis	Questionnaires/Interviews	Crowdsourcing platforms
Twitter	Kwon et al. [111] (R), Zubiaga et al. [211] (R), Thomson et al. [174] (R)	Morris et al. [130] (CA)	Ozturk et al. [136] (R), McCreadie et al. [123] (R)
Facebook	Zollo et al. [209] (CT), Zollo et al. [208] (CT), Bessi et al. [37] (CT)	Marchi [121] (B)	X
Other	Dang et al. [61] (R)	Chen et al. [55] (F), Kim and Bock [102] (R), Feldman [68] (B), Brewer et al. [44] (S) Winerburg and McGrew [190] (CA)	X

**Table 1:** Studies of user perception and interaction with false information on OSNs. The table depicts the main methodology of each paper as well as the considered OSN (if any). Also, where applicable, we report the type of false information that is considered (see bold markers and cf. with Section 2.1).

- **Influence.** This motive refers to the intent of misleading other people in order to influence their decisions, or manipulate public opinion with respect to specific topics. This motive can be distinguished into two general categories; 1) aiming to get leverage or followers (*power*) and 2) changing the norms of the public by disseminating false information. This is particularly worrisome on political matters [132], where individuals share false information to enhance an individuals’ public image or to hurt the public image of opposing politicians, especially during election periods.
- **Sow Discord.** In specific time periods individuals or organizations share false information to sow confusion or discord to the public. Such practices can assist in pushing a particular entity’s agenda; we have seen some examples on the political stage where foreign governments try to seed confusion in another country’s public for their own agenda [178].
- **Profit.** Many actors in the false information ecosystem seek popularity and monetary profit for their organization or website. To achieve this, they usually disseminate false information that increases the traffic on their website. This leads to increased ad revenue that results in monetary profit for the organization or website, at the expense of manipulated users. Some examples include the use of clickbait techniques, as well as fabricated news to increase views of articles from fake news sites that are disseminated via OSNs [143, 26]
- **Passion.** A considerable amount of users are passionate about a specific idea, organization, or entity. This affects their judgment and can contribute to the dissemination of false information. Specifically, passionate users are blinded by their ideology and perceive the false information as correct, and contribute in its overall propagation [85].
- **Fun.** As discussed in the previous section, online trolls are usually diffusing false information for their amusement. Their actions can sometimes inflict considerable damage to other individuals (e.g., see Doxing [165]), and thus should not be taken lightly.

Again, similarly to Sections 2.1 and 2.2, we have overlap among the presented motives. For instance, a political person may disseminate false information for political influence and because he is passionate about a specific idea.

### 3 User Perception of False Information

In this section, we describe work that study how users perceive and interact with false information on OSNs. Existing work use the following methodologies in understanding how false information is perceived by users: (i) by analyzing large-scale datasets obtained from OSNs; and (ii) by receiving input from users either from questionnaires, interviews, or through crowdsourcing marketplaces (e.g., Amazon Mechanical Turk, AMT [1]). Table 1 summarizes the studies on user perception, as well as their methodology and the considered OSN. Furthermore, we annotate each entry in Table 1 with the type of false information that each work considers. The remainder of this section provides an overview of the studies on understanding users’ perceptions on false information.

#### 3.1 OSN data analysis

Previous work focuses on extracting meaningful insights by analyzing data obtained from OSNs. From Table 1 we observe that previous work, leverages data analysis techniques to mainly study how users perceive and interact with rumors and conspiracy theories.

**Rumors.** Kwon et al. [111] study the propagation of rumors in Twitter, while considering findings from social and psychological studies. By analyzing 1.7B tweets, obtained from [52], they find that: 1) users that spread rumors and non-rumors have similar registration age and number of followers; 2) rumors have a clearly different writing style; 3) sentiment in news depends on the topic and not on the credibility of the post; and 4) words related to social relationships are more frequently used in rumors. Zubiaga et al. [211] analyze 4k tweets related to rumors by using journalists to annotate rumors in real time. Their findings indicate that true rumors resolved faster than false rumors and that the general tendency for users is to support every unverified rumor. However, the latter is less prevalent to reputable user accounts (e.g., reputable news outlets) that usually share information with evidence. Thomson et al. [174] study Twitter’s activity regarding the Fukushima Daiichi nuclear power plant disaster in Japan. The authors undertake a categorization of the messages according to their user, location, language, type, and credibility of the source. They observe that anonymous users, as well as users that live far away from the disaster share more information from less credible sources. Finally, Dang et al. [61] focus on the users that interact with rumors on Reddit by studying a popular false rumor (i.e., Obama is a Muslim). Specifically, they distinguish users into three main categories: the ones that support false rumors, the ones that refute false rumors and the ones that joke about a rumor. To identify these users they built a Naive Bayes classifier that achieves an accuracy of 80% and find that more than half of the users joked about this rumor, 25% refuted the joke and only 5% supported this rumor.

**Conspiracy Theories.** Zollo et al. [209] study the emotional dynamics around conversations regarding science and conspiracy theories. They do so by collecting posts from 280k users on Facebook pages that post either science or conspiracy theories posts. Subsequently, they use Support Vector Machines (SVMs) to identify the sentiment values of the posts, finding that sentiment is more negative on pages with conspiracy theories. Furthermore, they report that as conversations grow larger, the overall negative sentiment in the comments increases. In another work, Zollo et al. [208] perform a quantitative analysis of 54M Facebook users, finding the existence of well-formed communities for the users that interact with science and conspiracy news. They note that users of each community interact within the community and rarely outside of it. Also, debunking posts are rather inefficient and user exposure to such content increases the overall interest in conspiracy theory posts. Similarly, Bessi et al. [37] study how conspiracy theories and news articles are consumed on Facebook, finding that polarized users contribute more in the diffusion of conspiracy theories, whereas this does not apply for news and their respective polarized users.

### 3.2 Questionnaires/Interviews

To get insights on how users perceive the various types of false information, some of the previous work conducted questionnaires or interviews. The majority of the work aims to understand how younger users (students or teenagers) interact and perceive false information.

**Credibility Assessment.** Morris et al. [130] highlight that users are influenced by several features related to the author of a tweet like their Twitter username when assessing the credibility of information. Winerburg and McGrew [190] study whether users with different backgrounds have differences in their credibility assessments. To achieve this they conducted experiments with historians, fact-checkers, and undergraduate students, finding that historians and students can easily get manipulated by official-looking logos and domain names.

**Biased.** Marchi [121] focus on how teenagers interact with news on Facebook by conducting interviews with 61 racially diverse teenagers. The main findings of this study is that teenagers are not very interested in consuming news (despite the fact that their parents do) and that they demonstrate a preference to news that are opinionated when compared to objective news. Similarly, Feldman [68] focus on biased news and conduct 3 different studies with the participants randomly exposed to 2 biased and 1 non-biased news. The participants were asked to provide information about the news that allowed the authors to understand the perceived bias. They find that participants are capable of distinguishing bias in news articles; however, participants perceived lower bias in news that agree with their ideology/viewpoints.

**Fabricated.** Chen et al. [55] use questionnaires on students from Singapore with the goal to unveil the reasons that users with no malicious intent share false information on OSNs. They highlight that female students are more prone in sharing false information, and that students are willing to share information of any credibility just to initiate conversations or because the content seems interesting.

**Rumors.** Kim and Bock [102] study the rumor spreading behavior in OSNs from a psychological point of view by undertaking questionnaires on Korean students. They find that users’ beliefs results in either positive or negative emotion for the rumor, which affects the attitude and behavior of the users towards the rumor spreading.

**Satire.** Brewer et al. [44] indicate that satirical news programs can affect users’ opinion and political trust, while at the same time users tend to have stronger opinion on matters that they have previously seen in satirical programs.

### 3.3 Crowdsourcing platforms

Other related work leverages crowdsourcing platform to get feedback from users about false information. We note that, to the best of our knowledge, previous work that used crowdsourcing platforms focused on rumors.

**Rumors.** Ozturk et al. [136] study how users perceive health-related rumors and if they are willing to share them on Twitter. For acquiring the rumors, they crawl known health-related websites such as Discovery, Food Networks and National Institute of Health websites. To study the user perceptions regarding these rumors, they use AMT where they query 259 participants about ten handpicked health-related rumors. The participants were asked whether they will share a specific rumor or a message that refutes a rumor or a rumor that had a warning on it (i.e., “this message appeared in a rumor website”). Their results indicate that users are less likely to share a rumor that is accompanied with a warning or a message that refutes a rumor. Through simulations, they demonstrate that this approach can help in mitigating the spread of rumors on Twitter. Finally, McCreadie et al. [123] use crowdsourcing on three Twitter datasets related to emergency situations during 2014, in order to record users’ identification of rumors. Their results note that users were able to label most of the tweets correctly, while they note that tweets that contain controversial information are harder to distinguish.

### 3.4 User Perception - Future Directions

The studies discussed in this section aim to shed light on how users *perceive* false information on the Web. Overall the main take-away points from the reviewed related work are: 1) teenagers are not interested in consuming news; 2) students share information of any credibility just to initiate conversations; 3) in most cases, adults can identify bias in news and this task is harder when the news are biased towards the reader’s ideology; and 4) users can mostly identify rumors except the ones that contain controversial information. After reviewing the literature, we identify a few gaps in our understanding of false information perception. First, there is a lack of rigorous temporal analysis of user perception around the dissemination of false information and/or conspiracy theories. For example, perceptions might differ during the course of evolution of any particular conspiracy theory. Next, none of the studies reviewed take into consideration the interplay between *multiple* OSNs. Users on one platform might perceive the same false information differently depending on a variety of factors. For example, certain communities might be focused around a particular topic, affecting their susceptibility to false information on that topic, or the way the platform calls home presents information (e.g., news feed) can potentially influence how it users perceive false information. This issue becomes further muddled when considering users that are active on multiple platforms. In particular, we note that there is a substantial gap regarding *which* OSNs have been studied with respect to false information; e.g., YouTube, which has become a key player in information consumption. Finally, we would like to note that the European Union has recently approved a new regulation with regard to data protection, known as General Data Protection Regulation (GDPR) [65]. Therefore, we strongly advise researchers working in this line of work to study the new regulation and make sure that users give their explicit consent for participating in studies that aim to understand how users perceive false information.

## 4 Propagation of False Information

Understanding the dynamics of false information is of paramount importance as it gives useful insights regarding the problem. Table 2 summarizes the studies of false information propagation at OSNs, their methodology, as well as the corresponding type of false information according to the typology in Section 2.1. The research community focuses on studying the propagation by either employing data analysis techniques or mathematical and statistical approaches. Furthermore, we note the efforts done on providing systems that visualize the propagation dynamics of false information. Below, we describe the studies that are mentioned in Table 2 by dedicating a subsection for each type of methodology.

### 4.1 OSN Data Analysis

**Rumors.** Mendoza et al. [125] study the dissemination of false rumors and confirmed news on Twitter the days following the 2010 earthquake in Chile. They analyze the propagation of tweets for confirmed news and for rumors finding that the propagation of rumors differs from the confirmed news and that an aggregate analysis on the tweets can distinguish the rumors from the confirmed news. Similarly, Starbird et al. [168] study rumors regarding the 2013 Boston Bombings on Twitter and confirm both findings from Mendoza et al. [125]. In a similar notion, Nadamoto et al. [131] analyze the behavior of the Twitter community during disasters (Great East Japan Earthquake in 2011) when compared to a normal time period; finding that the spread of rumors during a disaster situation is different from a normal situation. That is in disaster situations, the hierarchy of tweets is shallow whereas in normal situations the tweets follow a deep hierarchy.

Others focused on understanding how rumors can be controlled and shed light on which types of accounts can help stop the rumor spread. Oh et al. [135] study Twitter data about the 2010 Haiti Earthquake and find that credible sources contribute in rumor controlling, while Andrews et al. [34] find that official accounts can contribute in stopping the rumor propagation by actively engaging in conversations related to the rumors.

Arif et al. [35] focus on the 2014 hostage crisis in Sydney. Their analysis include three main perspectives; (i) volume (i.e., number of rumor-related messages per time interval); (ii) exposure (i.e., number of individuals that were exposed to the rumor) and (iii) content production (i.e., if the content is written by the particular user or if it is a share). Their results highlight all three perspectives are important in understanding the dynamics of rumor propagation. Friggeri et al. [72] use known rumors that are

Platform	OSN data analysis	Epidemic & Statistical Modeling	Systems
Twitter	Mendoza et al. [125] (R), Oh et al. [135] (R), Andrews et al. [34] (R), Gupta et al. [77] (F), Starbird et al. [168] (R), Arif et al. [35] (R), Situngkir [162] (H), Nadamoto et al. [131] (R), Vosoughi et al. [181] (F)	Jin et al. [94] (R), Doerr et al. [63] (R), Jin et al. [95] (R)	Finn et al. [71] (R), Shao et al. [160] (F)
Facebook	Friggeri et al. [72] (R), Del Vicario et al. [62] (CT), Anagnostopoulos et al. [32] (CT)	Bessi [36] (CT)	X
Other	Ma and Li [119] (R), Zannettou et al. [200] (B)	Shah et al. [158] (R), Seo et al. [157] (R), Wang et al. [186] (R)	Dang et al. [60] (R)
Sina Weibo	X	Nguyen et al. [133] (R)	X

**Table 2:** Studies the focus on the propagation of false information on OSNs. The table summarizes the main methodology of each paper as well as the considered OSNs. Also, we report the type of false information that is considered (see bold markers and cf. with Section 2.1)

obtained through Snopes [20], a popular site that covers rumors, to study the propagation of rumors on Facebook. Their analysis indicates that rumors’ popularity is bursty and that a lot of rumors change over time, thus creating rumor variants. These variants aim to reach a higher popularity burst. Also, they note that rumors re-shares which had a comment containing a link to Snopes had a higher probability to be deleted by their users.

Finally, Ma and Li [119] study the rumor propagation process when considering a two-layer network; one layer is online (e.g., Twitter) and one layer is offline (e.g., face-to-face). Their simulations indicate that rumor spread is more prevalent in a two-layer network when compared with a single-layer offline network. The intuition is that in an offline network the spread is limited by the distance, whereas this constraint is eliminated in a two-layer network that has an online social network. Their evaluation indicates that in a two-layer network the spreading process on one layer does not affect the spreading process of the other layer; mainly because the interlayer transfer rate is less effective from an offline to an online network when compared with that from an OSN.

**Fabricated.** Gupta et al. [77] study the propagation of false information on Twitter regarding the 2013 Boston Marathon Bombings. To do so, they collect 7.9M unique tweets by using keywords about the event. Using real annotators, they annotate 6% of the whole corpus that represents the 20 most popular tweets during this crisis situation (i.e., the 20 tweets that got retweeted most times). Their analysis indicate that 29% of the tweets were false and a large number of those tweets were disseminated by reputable accounts. This finding contradicts with the findings of Oh et al. [135], which showed that credible accounts help stop the spread of false information, hence highlighting that reputable accounts can share bad information too. Furthermore, they note that out of the 32K accounts that were created during the crisis period, 19% of them were deleted or suspended by Twitter, indicating that accounts were created for the whole purpose of disseminating false information.

Vosoughi et al. [181] study the diffusion of false and true stories in Twitter over the course of 11 years. They find that false stories propagate faster, farther, and more broadly when compared to true stories. By comparing the types of false stories, they find that these effects were more intensive for political false stories when compared to other false stories (e.g., related to terrorism, science, urban legends, etc.).

**Hoaxes.** Situngkir [162] observe an empirical case in Indonesia to understand the spread of hoaxes on Twitter. Specifically, they focus on a case where a Twitter user with around 100 followers posted a question of whether a well-known individual is dead. Interestingly, the hoax had a large population spread within 2 hours of the initial post and it could be much larger if a popular mainstream medium did not publicly deny the hoax. Their findings indicate that a hoax can easily spread to the OSN if there is collaboration between the recipients of the hoax. Again, this work highlights, similarly to Oh et al. [135] that reputable accounts can help in mitigating the spread of false information.

**Conspiracy Theories.** Del Vicario et al. [62] analyze the cascade dynamics of users on Facebook when they are exposed to conspiracy theories and scientific articles. They analyze the content of 67 public pages on Facebook that disseminate conspiracy theories and science news. Their analysis indicates the formulation of two polarized and homogeneous communities for each type of information. Also, they note that despite the fact that both communities have similar content consumption patterns, they have different cascade dynamics. Anagnostopoulos et al. [32] study the role of homophily and polarization on the spread of false information by analyzing 1.2M Facebook users that interacted with science and conspiracy theories. Their findings indicate that user’s interactions with the articles correlate with the interactions of their friends (homophily) and that frequent exposure to conspiracy theories (polarization) determines how viral the false information is in the OSN.



**Biased.** Zannettou et al. [200], motivated by the fact that the information ecosystem consists of multiple Web communities, study the propagation of news across multiple Web communities. To achieve this, they study URLs from 99 mainstream and alternative news sources on three popular Web communities: Reddit, Twitter, and 4chan. Furthermore, they set out to measure the influence that each Web community has to each other, using a statistical model called Hawkes Processes. Their findings indicate that small fringe communities within Reddit and 4chan have a substantial influence to mainstream OSNs like Twitter.

## 4.2 Epidemic and Statistical Modeling

**Rumors.** Jin et al. [94] use epidemiological models to characterize cascades of news and rumors in Twitter. Specifically, they use the SEIZ model [38] which divides the user population in four different classes based on their status; (i) Susceptible; (ii) Exposed; (iii) Infected and (iv) Skeptic. Their evaluation indicates that the SEIZ model is better than other models and it can be used to distinguish rumors from news in Twitter. In their subsequent work, Jin et al. [95] perform a quantitative analysis on Twitter during the Ebola crisis in 2014. By leveraging the SEIZ model, they show that rumors spread in Twitter the same way as legitimate news.

Doerr et al. [63] use a mathematical approach to prove that rumors spread fast in OSNs (similar finding with Vosoughi et al. [181]). For their simulations they used real networks that represent the Twitter and Orkut Social Networks topologies obtained from [52] and SNAP [19], respectively. Intuitively, rumors spread fast because of the combinations of few large-degree nodes and a large number of small-degree nodes. That is, small-degree nodes learn a rumor once one of their adjacent nodes knows it, and then quickly forward the rumor to all adjacent nodes. Also, the propagation allows the diffusion of rumors between 2 large-degree nodes, thus the rapid spread of the rumor in the network.

Several related work focus on finding the source of the rumor. Specifically, Shah et al. [158] focus on detecting the source of the rumor in a network by defining a new rumor spreading model and by forming the problem as a maximum likelihood estimation problem. Furthermore, they introduce a new metric, called *rumor centrality*, which essentially specifies the likelihood that a particular node is the source of the rumor. This metric is evaluated for all nodes in the network by using a simple linear time message-passing algorithm, hence the source of the rumor can be found by selecting the node with the highest rumor centrality. In their evaluation, they used synthetic small-world and scale-free real networks to apply their rumor spreading model and they show that they can distinguish the source of a rumor with a maximum error of 7-hops for general networks, and with a maximum error of 4-hops for tree networks. Seo et al. [157] aim to tackle the same problem by injecting monitoring nodes on the social graph. They propose an algorithm that considers the information received by the monitoring nodes to identify the source. They indicate that with sufficient number of monitoring nodes they can recognize the source with high accuracy. Wang et al. [186] aim to tackle the problem from a statistical point of view. They propose a new detection framework based on rumor centrality, which is able to support multiple snapshots of the network during the rumor spreading. Their evaluation based on small-world and scale-free real networks note that by using two snapshots of the network, instead of one, can improve the source detection. Finally, Nguyen et al. [133] aim to find the  $k$  most suspected users where a rumor originates by proposing the use of a reverse diffusion process in conjunction with a ranking process.

**Conspiracy Theories.** Bessi [36] perform a statistical analysis of a large corpus (354k posts) of conspiracy theories obtained from Facebook pages. Their analysis is based on the Extreme Value Theory branch of statistics [5] and they find that extremely viral posts (greater than 250k shares) follow a Poisson distribution.

## 4.3 Systems

**Rumors.** Finn et al. [71] propose a web-based tool, called TwitterTrails, which enables users to study the propagation of rumors in Twitter. TwitterTrails demonstrates indications for bursty activity, temporal characteristics of propagation, and visualizations of the re-tweet networks. Furthermore, it offers advanced metrics for rumors such as level of visibility and community's skepticism towards the rumor (based on the theory of h-index [10]). Similarly, Dang et al. [60] propose RumourFlow, which visualizes rumors propagation by adopting modeling and visualization tools. It encompasses various analytical tools like semantic analysis and similarity to assist the user in getting a holistic view of the rumor spreading and its various aspects. To demonstrate their system, they collect rumors from Snopes and conversations from Reddit.

**Fabricated.** Shao et al. [160] propose Hoaxy, a platform that provides information about the dynamics of false information propagation on Twitter as well as the respective fact checking efforts.

## 4.4 Propagation of False Information - Future Directions

In this section, we provided an overview of the existing work that focuses on the propagation of false information on the Web. Some of the main take-aways from the literature review on the propagation of false information are: 1) Accounts on social networks are created with the sole purpose of disseminating false information; 2) False information is more persistent than corrections; 3) The popularity of false information follow a bursty activity; 4) Users on Web communities create polarized communities that disseminate false information; 5) Reputable or credible accounts are usually useful in stopping the spread of

Platform	Machine Learning	Systems	Other models/algorithms
Twitter	Castillo et al. [51] (CA), Gupta and Kumaraguru [75] (CA), Kwon et al. [112] (R), Yang et al. [198] (R), Liu et al. [116] (R), Wu et al. [194] (R), Gupta et al. [76] (CA), AlRubaian et al. [30] (CA), Hamidian and Diab [79] (R), Giasemidis et al. [73] (R), Kwon et al. [110] (R), Volkova et al. [179] (CA)	Resnick et al. [151] (R), Vosoughi et al. [180] (R), Jaho et al. [93] (CA)	Qazvinian et al. [147] (R) (rumor retrieval model), Zhao et al. [205] (R) (clustering), Farajtabar et al. [67] (F) (hawkes process), Kumar and Geethakumari [107] (F) (algorithm with psychological cues)
Sina Weibo	Yang et al. [195] (R), Wu et al. [193] (R), Liang et al. [115] (R), Zhang et al. [204] (R),	Zhou et al. [207] (CA)	X
Twitter and Sina Weibo	Ma et al. [118] (CA) Ma et al. [117] (R)	X	Jin et al. [99] (CA) (graph optimization)
Facebook	Tacchini et al. [170] (H), Conti et al. [59] (CT)	X	X
Wikipedia and/or other articles	Qin et al. [148] (R), Rubin et al. [155] (S), Kumar et al. [109] (H), Chen et al. [56] (CL), Chakraborty et al. [53] (CL), Potthast et al. [145] (CL), Biyani et al. [39] (CL), Wang [185] (F), Anand et al. [33] (CL)	X	Potthast et al. [144] (B) (unmasking)
Other	Afroz et al. [27] (H), Maigrot et al. [120] (H), Zannettou et al. [202] (CL)	Vukovic et al. [182] (H)	Jin et al. [97] (CA) (hierarchical propagation model), Chen et al. [57] (H) (Levenshtein Distance)

**Table 3:** Studies that focus on the detection of false information on OSNs. The table demonstrates the main methodology of each study, as well as the considered OSNs. Also, we report the type of false information that is considered (see bold markers and cf. with Section 2.1, CA corresponds to Credibility Assessment and refers to work that aim to assess the credibility of information).

false information; however we need to pay particular attention as previous work (see Gupta et al. [77]) has showed that they also share false information; 6) Being able to detect the source of false information is a first step towards stopping the spread of false information on Web communities and several approaches exist that offer acceptable performance. As future directions to this line of work, we propose studying the problem from a multi-platform point of view. That is, study how information propagates across *multiple communities* and fusing information that exists in *multiple formats* (e.g., images, textual claims, URLs, video, etc.). Furthermore, systems or tools that visualize the propagation of information across OSNs do not exist. These type of tools will enable a better understanding of false information propagation, as well as finding the source of information. Finally, to the best of our knowledge, the propagation of information via *orchestrated campaigns* has not been rigorously studied by the research community. An example of such a campaign is the posting comments in YouTube video by users of 4chan [86].

## 5 Detection and Containment of False Information

### 5.1 Detection of false information

Detecting false information is not a straightforward task, as it appears in various forms, as discussed in Section 2. Table 3 summarizes the studies that aim to solve the false information detection problem, as well as their considered OSNs and their methodology. Most studies try to solve the problem using handcrafted features and conventional machine learning techniques. Recently, to avoid using handcrafted features, the research community used neural networks to solve the problem (i.e., Deep Learning techniques). Furthermore, we report some systems that aim to inform users about detected false information. Finally, we also note a variety of techniques that are proposed for the detection and containment of false information, such as epidemiological models, multivariate Hawkes processes, and clustering. Below, we provide more details about existing work grouped by methodology and the type of information, according to Table 3.

### 5.1.1 Machine Learning

**Credibility Assessment.** Previous work leverage machine learning techniques to assess the credibility of information. Specifically, Castillo et al. [51] analyze 2.5k trending topics from Twitter during 2010 to determine the credibility of information. For labeling their data they utilize crowdsourcing tools, namely AMT, and propose the use of conventional machine learning techniques (SVM, Decision Trees, Decision Rules, and Bayes Networks) that take into account message-based, user-based, topic-based and propagation-based features. Gupta and Kumaraguru [75] analyze tweets about fourteen high impact news events during 2011. They propose the use of supervised machine learning techniques with a relevance feedback approach that aims to rank the tweets according to their credibility score. AlRubaian et al. [30] propose the use of a multi-stage credibility assessment platform that consists of a relative importance component, a classification component, and an opinion mining component. The relative importance component requires human experts and its main objective is to rank the features according to their importance. The classification component is based on a Naive Bayes classifier, which is responsible for classifying tweets by taking the output of the relative importance component (ranked features), while the opinion mining component captures the sentiment of the users that interact with the tweets. The output of the three components is then combined to calculate an overall assessment. Ma et al. [118] observe that typically the features of messages in microblogs vary over time and propose the use of an SVM classifier that is able to consider the messages features in conjunction with how they vary over time. Their experimental evaluation, based on Twitter data provided by [51] and on a Sina Weibo dataset, indicate that the inclusion of the time-varying features increase the performance between 3% and 10%.

All of the aforementioned work propose the use of supervised machine learning techniques. In contrast, Gupta et al. [76] propose a semi-supervised model that ranks tweets according to their credibility in real-time. For training their model, they collect 10M tweets from six incidents during 2013, while they leverage CrowdFlower [2] to obtain groundtruth. Their system also includes a browser extension that was used by approx. 1.1k users in a 3-month timespan, hence computing the credibility score of 5.4M tweets. Their evaluation indicates that 99% of the users were able to receive credibility scores under 6 seconds. However, feedback from users for approx. 1.2k tweets indicate that 60% of the users disagreed with the predicted score.

Volkova et al. [179] motivated by the performance gains of deep learning techniques, propose the use of neural networks to distinguish news into satire, hoaxes, clickbait, and propaganda news. They collect 130k news posts from Twitter and propose the use of neural networks that use linguistic and network features. Their findings indicate that Recurrent and Convolutional neural networks exhibit strong performance in distinguishing news in the aforementioned categories.

**Rumors.** Kwon et al. [112] propose the use of Decision Trees, Random Forest, and SVM for detecting rumors on Twitter. Their models leverage temporal, linguistics, and structural features from tweets and can achieve precision and recall scores between 87% and 92%. Yang et al. [198] propose the use of a hot topic detection mechanism that work in synergy with conventional machine learning techniques (Naive Bayes, Logistic Regression and Random Forest). Liu et al. [116] demonstrate the feasibility of a real-time rumoring detection system on Twitter. To achieve real-time debunking of rumors, they propose the use of an SVM classifier that uses beliefs from the users in conjunction with traditional rumor features from [51, 195]. Their evaluation demonstrates that for new rumors (5-400 tweets), the proposed classifier can outperform the models from [51, 195]. Furthermore, they compare their approach with human-based rumor debunking services (Snopes and Emergent), showing that they can debunk 75% of the rumors earlier than the corresponding services. Similarly, Kwon et al. [110] study the rumor classification task with a particular focus on the temporal aspect of the problem, by studying the task over varying time windows on Twitter. By considering user, structural, linguistic, and temporal features, they highlight that depending on the time window, different characteristics are more important than others. For example, at early stages of the rumor propagation, temporal and structural are not available. To this end, they propose a rumor classification algorithm that achieves satisfactory accuracy both on short and long time windows.

Hamidian and Diab [79] propose a supervised model that is based on the Tweet Latent Vector (TLV), which is an 100-dimensional vector, proposed by the authors, that encapsulates the semantics behind a particular tweet. For the classification task, they use an SVM Tree Kernel model that achieves 97% on two Twitter datasets. Giasemidis et al. [73] study 72 rumors in Twitter by identifying 80 features for classifying false and true rumors. These features include diffusion and temporal dynamics, linguistics, as well as user-related features. For classifying tweets, they use several machine learning techniques and conclude that Decision Trees achieve the best performance with an accuracy of 96%. Yang et al. [195] study the rumor detection problem in the Sina Weibo OSN. For the automatic classification task of the posts they use SVMs that take as input various features ranging from content-based to user- and location-based features. Their evaluation shows that the classifier achieves an accuracy of approximately 78%. Similarly to the aforementioned work, Wu et al. [193] try to tackle the rumor detection problem in the Sina Weibo OSN by leveraging SVMs. Specifically, they propose an SVM classifier which is able to combine a normal radial basis function, which captures high level semantic features, and a random walk graph kernel, which captures the similarities between propagation trees. These trees encompass various details such as temporal behavior, sentiment of re-posts, and user details. Liang et al. [115] study the problem of rumor detection using machine learning solutions that take into account users' behavior in the Sina Weibo OSN. Specifically, they introduce 3 new features that are shown to provide up to 20% improvement when compared with baselines. These features are: 1) average number of followees per day; 2) average number of posts per day; and 3) number of possible microblog sources. Zhang et al. [204] propose various implicit features that can assist in the detection of rumors. Specifically, they evaluate an SVM classifier against the Sina Weibo dataset proposed in [195] with the

following features: 1) content-based implicit features (sentiment polarity, opinion on comments and content popularity); 2) user-based implicit features (influence of user to network, opinion re-tweet influence, and match degree of messages) and 3) shallow message features that are proposed by the literature. Their evaluation shows that the proposed sets of features can improve the precision and recall of the system by 7.1% and 6.3%, respectively. Qin et al. [148] propose the use of a new set of features for detecting rumors that aim to increase the detection accuracy; namely novelty-based and pseudo-feedback features. The novelty-based features consider reliable news to find how similar is a particular rumor with reliable stories. The pseudo-feedback features take into account information from historical confirmed rumors to find similarities. To evaluate their approach, they obtain messages from the Sina Weibo OSN and news articles from Xinhua News Agency [25]. They compare an SVM classifier, which encompasses the aforementioned set of features and a set of other features (proposed by the literature), with the approaches proposed by [195, 116]. Their findings indicate that their approach provides an improvement between 17% and 20% in terms of accuracy. Similarly to [148], Wu et al. [194] propose a system that uses historical data about rumors for the detection task. Their system consists of a feature selection module, which categorizes and selects features, and a classifier. For constructing their dataset they use Snopes and the Twitter API to retrieve relevant tweets, acquiring in total 10k tweets, which are manually verified by annotators. In their evaluation, they compare their system with various baselines finding that the proposed system offers enhanced performance in rumor detection with an increase of 12%-24% for precision, recall, and F1-score metrics. Ma et al. [117] leverage Recurrent neural networks to solve the problem of rumor detection in OSNs. Such techniques are able to learn hidden representations of the input without the need for hand-crafted features. For evaluating their model, they construct two datasets; one from Twitter and one from Sina Weibo. For the labeling of their messages they use Snopes for Twitter and the official rumor-busting service of Sina Weibo's OSN. Their evaluation shows an accuracy of 91% on the Sina Weibo dataset and 88% on the Twitter dataset.

**Hoaxes.** Tacchini et al. [170] study hoaxes in Facebook and argue that they can accurately discern hoax from non-hoax posts by simply looking at the users that liked the posts. Specifically, they propose the use of Logistic Regression that classifies posts with features based on users' interactions. Their evaluation demonstrate that they can identify hoaxes with an accuracy of 99%. Kumar et al. [109] study the presence of hoaxes in Wikipedia articles by considering 20k hoax articles that are explicitly flagged by Wikipedia editors. They find that most hoaxes are detected quickly and have little impact, however, a small portion of these hoaxes have a significant life-span and are referenced a lot across the Web. By comparing the "successful" hoaxes with failed hoaxes and legitimate articles, the authors highlight that the successful hoaxes have notable differences in terms of structure and content. To this end, they propose the use of a Random Forest classifier to distinguish if articles are hoaxes. Their evaluation reports that their approach achieves an accuracy of 92% and that is able to outperform human judgments by a significant margin (20%). Maigrot et al. [120] propose the use of a multi-modal hoax detection system that fuses the diverse modalities pertaining to a hoax. Specifically, they take into consideration the text, the source, and the image of tweets. They observe higher performance when using only the source or text modality instead of the combination of all modalities.

**Conspiracy Theories.** Conti et al. [59] focus on identifying conspiracy theories in OSNs by considering only the structural features of the information cascade. The rationale is that such features are difficult to be tampered by malicious users, which aim to avoid detection from classification systems. For their dataset they use data from [37], which consist of scientific articles and conspiracy theories. For classifying their Facebook data they propose conventional machine learning techniques and they find that it is hard to distinguish a conspiracy theory from a scientific article by only looking at their structural dynamics (F1 -score not exceeding 65%).

**Satire.** Rubin et al. [155] propose the use of satirical cues for the detection of false information on news articles. Specifically, they propose the use of five new set of features, namely absurdity, humor, grammar, negative affect, and punctuation. Their evaluation shows that by using an SVM algorithm with the aforementioned set of features and others proposed by the literature, they can detect satirical news with 90% precision and 84% recall.

**Clickbait.** Several studies focus on the detection of clickbait on the Web using machine learning techniques. Specifically, Chen et al. [56] propose tackling the problem using SVMs and Naive Bayes. Also, Chakraborty et al. [53] propose the use of SVM and a browser add-on to offer a system to users for news articles. Potthast et al. [145] proposes the use of Random Forest for detecting clickbait tweets. Moreover, Biyani et al. [39] propose the use of Gradient Boosted Decision Trees for clickbait detection in news articles and show that the degree of informality in the content of the landing page can help in finding clickbait articles. Anand et al. [33] is the first work that suggests the use of deep learning techniques for mitigating the clickbait problem. Specifically, they propose the use of Recurrent Neural Networks in conjunction with word2vec embeddings [127] for identifying clickbait news articles. Similarly, Zannettou et al. [202] use deep learning techniques to detect clickbaits on YouTube. Specifically, they propose a semi-supervised model based on variational autoencoders (deep learning). Their evaluation indicates that they can detect clickbaits with satisfactory performance and that YouTube's recommendation engine does not consider clickbait videos in its recommendations.

**Fabricated.** Wang [185] presents a dataset that consists of 12.8k manually annotated short statements obtained from PolitiFact. They propose the use of Convolutional neural networks for fusing linguistic features with metadata (e.g., who is the author of the statement). Their evaluation demonstrates that the proposed model outperforms SVM and Logistic Regression algorithms.

### 5.1.2 Systems

**Rumors.** Resnick et al. [151] propose a system called RumorLens, which aims to discover rumors in a timely manner, provide insights regarding the rumor’s validity, and visualize a rumor’s propagation. To achieve the aforementioned, RumorLens leverages data mining techniques alongside with a visual analysis tool. However, their system raises scalability issues as it highly depends on users’ labor, which provide labeling of tweets that are subsequently used for classifying tweets related to a particular rumor. Vosoughi et al. [180] propose a human-machine collaborative system that aims to identify rumors by disposing irrelevant data and ranking the relevant data. Their system consists of two components; the assertion detector and the hierarchical clustering module. The assertion detector is a classifier that uses semantic and syntactic features to find tweets that contain assertions. These tweets are then presented to the clustering module, which clusters the tweets according to the similarity of the assertions. During their evaluation, the authors state that for a particular incident (Boston Marathon Bombings) from a dataset of 20M tweets, their system managed to discard 50% of them using the assertion detector. Furthermore, the 10M relevant tweets are clustered somewhere between 100 and 1000 clusters, something that enables users to quickly search and find useful information easier.

**Credibility Assessment.** Jaho et al. [93] undertake a statistical analysis by crawling Twitter for 3 months and retrieve a dataset that includes 10M users. They propose a system that is based on contributor-related features (e.g., reputation, influence of source, etc.), content features (e.g., popularity, authenticity, etc.) and context features (e.g., coherence, cross-checking, etc.). Their system combines all the features and outputs a single metric that corresponds to the truthfulness of the message. Zhou et al. [207] note that calculating credibility in the granularity of message is not scalable, therefore they propose the calculation of credibility score per event. To this end, they propose a system that is able to collect related data from Sina Weibo using keywords and detect the credibility of a particular event. The credibility score is calculated by the combination of 3 sub-models; the user model, the propagation model, and the content model. Each one of the sub-models considers one aspect of the news and the overall score is calculated using weighted combination. The system is trained on a dataset that contains 73 real news and 73 fake news from approximately 50k posts. Their evaluation shows that the proposed system provides an accuracy close to 80% and that credibility scores are calculated within 35 seconds.

**Hoaxes.** Vukovic et al. [182] focus on hoaxes and propose the use of a detection system for email. The proposed system consists of a feed-forward neural network and a self-organizing map (SOM) and it is trained on a corpus of 298 hoax and 1370 regular emails. The system achieves an accuracy of 73% with a ratio of false positives equal to 4.9%. Afroz et al. [27] focus on detecting hoaxes by observing changes in writing style. The intuition is that people use different linguistic features when they try to obfuscate or change information from users. To detect hoaxes they propose the use of an SVM classifier that takes into account the following set of features: 1) lexical features; 2) syntactic features; 3) content features and 4) lying detection features obtained from [49, 80]. Their evaluation on various datasets indicates that the proposed system can detect hoaxes with an accuracy of 96%.

### 5.1.3 Other models/algorithms

**Rumors.** Qazvinian et al. [147] study the rumor detection problem on Twitter by retrieving tweets regarding rumors and leveraging manual inspectors to annotate it. Specifically, the annotators were asked whether tweets contained rumors or not and whether a user endorsed, debunked or was neutral about the rumors. The resulted dataset consists of approximately 10k annotated tweets and was analyzed to demonstrate the effectiveness of the following feature sets in identifying rumors: 1) content-based features; 2) network-based features and 3) Twitter-specific memes (hashtags and URLs). Furthermore, the paper proposes a rumor retrieval model that achieves 95% precision. Zhao et al. [205] are motivated by the fact that identifying false factual claims in each individual message is intractable. To overcome this, they adapt the problem in finding whole clusters of messages that their topic is a disputed factual claim. To do so, they search within posts to find specific phrases that are used from users who want to seek more information or to express their skepticism. For example, some enquiry phrases are ”Is this true?”, ”Really?” and ”What?”. Their approach uses statistical features of the clusters in order to rank them according to the likelihood of including a disputed claim. Their evaluations on real Twitter data indicate that among the top 50 ranked clusters, 30% of them are confirmed rumors.

**Fabricated.** Farajtabar et al. [67] propose a framework for tackling false information that combines a multivariate Hawkes process and reinforcement learning. Their evaluation highlights that their model shows promising performance in identifying false information in real-time on Twitter. Kumar and Geethakumari [107] measure the diffusion of false information by exploiting cues obtained from cognitive psychology. Specifically, they consider the consistency of the message, the coherency of the message, the credibility of the source, and the general acceptability of the content of the message. These cues are fused to an algorithm that aims to detect the spread of false information as soon as possible. Their analysis on Twitter reports that the proposed algorithm has a 90% True positive rate and a False positive rate less than 10%.

**Credibility Assessment.** Jin et al. [99] aim to provide verification of news by considering conflicting viewpoints on Twitter and Sina Weibo. To achieve this, they propose the use of a topic model method that identifies conflicting viewpoints. Subsequently they construct a credibility network with all the viewpoints and they formulate the problem as a graph optimization problem, which can be solved with an iterative approach. They compare their approach with baselines proposed in [51, 112], showing that their solution performs better. Jin et al. [97] propose a hierarchical propagation model to evaluate information credibility

in microblogs by detecting events, sub-events, and messages. This three-layer network assists in revealing vital information regarding information credibility. By forming the problem as a graph optimization problem, they propose an iterative algorithm, that boosts the accuracy by 6% when compared to an SVM classifier that takes into account only features obtained from the event-level network only.

**Biased.** Potthast et al. [144] study the writing style of hyperpartisan news (left-wing and right-wing) and mainstream news and how this style can be applied in hyperpartisan news detection. Their dataset consists of 1.6k news articles from three right-wing, three left-wings, and three mainstream news sites. For annotating the dataset they used journalists from BuzzFeed, who rated each article according to its truthfulness. By leveraging the Unmasking approach [105], the paper demonstrates that right-wing and left-wing hyperpartisan news exhibit similar writing style that differentiates from the mainstream news. To this end, they propose the use of Random Forest classifier that aims to distinguish hyperpartisanship. Their evaluation indicates that their style-based classifier can distinguish hyperpartisan news with an accuracy of 75%. However, when the same classifier is used to discern fake or real news, then the accuracy is 55%.

**Hoaxes.** Chen et al. [57] propose an email hoax detection system by incorporating a text matching method using the Levenshtein distance measure. Specifically, their system maintains a database of hoaxes that is used to calculate the distance between a potential hoax email and the stored hoaxes.

## 5.2 Containment of false information

Several studies focus on containing the diffusion of false information. Our literature review reveals that the majority of previous work on containment of rumors, while we also find one that focus on Hoaxes (see Tambuscio et al. [171]). Below we provide a brief overview of the studies that try to contain the spread of false information, while ensuring that the solutions are scalable.

**Rumors.** Tripathy et al. [177] propose a process, called "anti-rumor", which aims to mitigate the spreading of a rumor in a network. This process involves the dissemination of messages, which contradict with a rumor, from agents. The authors make the assumption that once a user receives an anti-rumor message, then he will never believe again the rumor, thus the spreading of a rumor is mitigated. Their evaluation, based on simulations, indicates the efficacy of the proposed approach. Budak et al. [45] formulate the problem of false information spreading as an optimization problem. Their aim is to identify a subset of the users that need to be convinced to spread legitimate messages in contrast with the bad ones that spread rumors. The paper shows that this problem is NP-hard and they propose a greedy solution as well as some heuristics to cope with scalability issues. Fan et al. [66] try to tackle the problem of false information propagation under the assumption that rumors originate from a particular community in the network. Similarly to other work, the paper tries to find a minimum set of individuals, which are neighbors with the rumor community to stop the rumor diffusion in the rest of the network. To achieve this, they propose the use of two greedy-based algorithms, which are evaluated in two real-world networks (Arxiv Hep and Enron). Their experimental results show that the proposed algorithms outperform simple heuristics in terms of the number of infected nodes in the network. However, as noted, the greedy algorithms are time consuming and are not applicable in large-scale networks. Kotnis et al. [106] propose a solution for stopping the spread of false information by training a set of individuals in a network that aim to distinguish and stop the propagation of rumors. This set of individuals is selected based on their degree in the network with the goal to minimize the overarching training costs. For evaluating their solution they create a synthetic network, which takes into account a calculated network degree distribution, based on [129]. Ping et al. [141] leverage Twitter data to demonstrate that sybils presence in OSNs can decrease the effectiveness of community-based rumor blocking approaches by 30%. To this end, they propose a Sybil-aware rumor blocking approach, which finds a subset of nodes to block by considering the network structure in conjunction with the probabilities of nodes being sybils. Their evaluation, via simulations on Twitter data, show that the proposed approach significantly decreases the number of affected nodes, when compared to existing approaches. He et al. [82] argue that existing false information containment approaches have different costs and efficiencies in different OSNs. To this end, they propose an optimization method that combines the spreading of anti-rumors and the block of rumors from influential users. The goal of their approach is to minimize the overarching cost of the method while containing the rumor within an expected deadline. To achieve this, they use the Pontryagin's maximum principle [104] on the Digg2009 dataset [87]. They find that spreading the truth plays a significant role at the start of the rumor propagation, whereas close to the deadline of containment the blocking of rumors approach should be used extensively. Huang et al. [91] aim to contain the false information spread by finding and decontaminating with good information, the smallest set of influential users in a network. To do so, they propose a greedy algorithm and a community-based heuristic, which takes into consideration the community structure of the underlying network. For evaluating their approach, they used traces from three networks; NetHEPT, NetHEPT\_WC and Facebook. Previous studies on false information containment [45, 134] assumed that when true and false information arrive the same time at a particular node, then the true information dominates. Wang et al. [184] state that the dominance of the information should be based on the influence of the neighbors in the network. With this problem formulation in mind, the paper proposes two approaches to find the smallest number of nodes that are required to stop the false information spread. Their evaluation is based on three networks obtained from Twitter, Friendster, and a random synthetic network. Evaluation comparisons with simple heuristics (random and

high degree) demonstrate the performance benefits of the proposed approaches. In a similar notion, Tong et al. [176] aim to increase performance motivated by the fact that greedy solutions, which include Monte Carlo simulations, are inefficient as they are computationally intensive. To overcome this, the paper proposes a random-based approach, which utilizes sampling with the aim to be both effective and efficient. The performance evaluations on real-world (obtained from Wikipedia and Epinions [4]) and synthetic networks demonstrate that the proposed solution can provide a 10x speed-up without compromising performance when compared to state-of-the-art approaches. Wang et al. [183] propose a model, called DRIMUX, which aims to minimize the influence of rumors by blocking a subset of nodes while considering users' experience. User experience is defined as a time threshold that a particular node is willing to wait while being blocked. Their model utilizes survival theory and takes into account global rumor popularity features, individual tendencies (how likely is a rumor to propagate between a pair of nodes) as well as the users' experience. Their evaluations on a Sina Weibo network, which consists of 23k nodes and 183k edges, indicate that the proposed model can reduce the overarching influence of false information.

**Hoaxes.** Tambuscio et al. [171] simulate the spread and debunking of hoaxes on networks. Specifically, they model the problem as a competition between believers (acknowledge the hoax) and fact checkers which reveal the hoax with a specific probability. To study their model they performed simulations on scale-free and random networks finding that a specific threshold for the probability of fact checkers exists and this indicates that the spread can be stopped with a specific number of fact checkers. However, the paper oversimplifies the problem by assuming all the nodes to have the same probability.

### 5.3 Detection and Containment of False Information - Future Directions

The main findings from the literature review of the detection and containment of false information are: 1) Machine learning techniques can assist in identifying false information. However, they heavily rely on handcrafted set of features and it is unclear if they generalize well on other datasets; 2) Containment of false information can be achieved by adding a set of good nodes that disseminate good information or information that refute false; and 3) The problem of detection of false information requires human-machine collaboration for effectively mitigating it.

Despite the fact that several studies exist attempting to detect and contain false information on the Web, the problem is still emerging and prevalent. This is mainly because the problem requires higher cognitive and context awareness that current systems do not have. To assist in achieving a better detection of false information on the Web, we foresee the following tangible research directions.

First, information on the Web exists in multiple formats, and thus false information is disseminated via textual claims, screenshots, videos, etc. Most studies, however, take into consideration only one format. To achieve a multi-format false information detection system requires correlating information in multiple formats, which in turn requires understanding the similarities and differences in the content each format delivers. To the best of our knowledge, a system that can meaningfully assist in detecting false information across multiple formats does not exist. Next, to the best of our knowledge, no prior work has rigorously assessed credibility based on user profiling. For example, a post from an expert on a particular subject should not be treated with the same weight as a post by a typical user. We foresee studying false information detection from a users' perspective is a way forward in effectively detecting and containing the spread of false information on the Web. Finally, most previous studies focus on detection and containment of false information on a single OSN, but the Web is much larger than any single platform or community. Therefore, future work should address the problem with a holistic view of the information ecosystem. This requires an understanding of how information jumps from one Web community to another, how Web communities influence each other, and how to correlate accounts that exist in multiple Web communities (e.g., how to find that a particular Facebook and Twitter account belong to the same user). Such an understanding will be particularly useful for containing the spread of false information from one Web community to another.

## 6 False Information in the political stage

Recently, after the 2016 US elections, the problem of false information dissemination got extensive interest from the community. Specifically, Facebook got openly accused for disseminating false information and that affected the outcome of the elections [6]. It is evident that dissemination of false information on the Web is used a lot for political influence. Therefore in this section we review the most relevant studies on the political stage. Table 4 reports the reviewed work as well as the main methodology and considered OSN.

### 6.1 Machine Learning

**Propaganda.** Ratkiewicz et al. [150] study political campaigns on Twitter that use multiple controlled accounts to disseminate support for an individual or opinion. They propose the use of a machine learning-based framework in order to detect the early stages of the spreading of political false information on Twitter. Specifically, they propose a framework that takes into consideration topological, content-based and crowdsourced features of the information diffusion in Twitter. Their experimental evaluation demonstrates that the proposed framework achieves more than 96% accuracy in the detection of political campaigns for data

Platform	Machine Learning	OSN Data Analysis	Other models/algorithms
Twitter	Ratkiewicz et al. [150] (P), Conover et al.[58] (P), Ferrara et al.[70] (P)	Wong et al. [191] (B), Golbeck and Hansen [74] (B), Jackson and Welles [92] (P), Hegelich and Janetzko[83] (P), Zannettou et al. [201] (P) Howard and Kollanyi[90] (P), Shin et al.[161] (R)	An et al. [31] (B) (distance model), Al-khateeb and Agarwal [28] (P) (social studies) Ranganath et al.[149] (P) (exhaustive search), Jin et al. [96] (R) (text similarity), Yang et al. [196] (B) (agenda-setting tool)
Digg	Zhou et al.[206] (B)	X	X
Sina Weibo	X	King et al. [103] (P), Yang et al. [197] (P)	X
News articles	Budak et al. [46] (B)	Woolley[192] (P)	X
Facebook	X	Allcot and Gentzkow[29] (P)	X

**Table 4:** Studies on the false information ecosystem on the political stage. The table demonstrates the main methodology of each study as well as the considered OSNs.

pertaining to the 2010 US midterm elections. Conover et al. [58] study Twitter on a six-week period leading to the 2010 US midterm elections and the interactions between right and left leaning communities. They leverage clustering algorithms and manually annotated data to create the re-tweets and mentions networks. Their findings indicate that the re-tweet network has limited connectivity between the right and left leaning communities, whereas this is not the case in the mentions networks. This is because, users try to inject different opinions on users with different ideologies, by using mentions on tweets, so that they change their stance towards a political individual or situation. Ferrara et al. [70] propose the use of a k-nearest neighbor algorithm with a dynamic warping classifier in order to capture promoted campaigns in Twitter. By extracting a variety of features (user-related, timing-related, content-related and sentiment-related features) from a large corpus of tweets they demonstrate that they can distinguish promoted campaigns with an AUC score close to 95% in a timely manner.

**Biased.** Zhou et al. [206] study Digg, a news aggregator site, and aim to classify users and articles to either liberal or conservative. To achieve this, they propose three semi-supervised propagation algorithms that classify users and articles based on users’ votes. The algorithms make use of a few labeled users and articles to predict a large corpus of unlabeled users and articles. The algorithms are based on the assumption that a liberal user is more likely to vote for a liberal article rather than a conservative article. Their evaluations demonstrate that the best algorithm achieves 99% and 96% accuracy on the dataset of users and articles, respectively. Budak et al. [46] use Logistic Regression to identify articles regarding politics from a large corpus of 803K articles obtained from 15 major US news outlets. Their algorithm filtered out 86% of the articles as non-political related, while a small subset of the remainder (approx. 11%) were presented to workers on AMT. The workers were asked to answer questions regarding the topic of the article, whether the article was descriptive or opinionated, the level of partisanship, and the level of bias towards democrats or republicans. Their empirical findings are that on these articles there are no clear indications of partisanship, some articles within the same outlet are left-leaning and some have right-leaning, hence reducing the overall outlet bias. Also, they note that usually bias in news articles is expressed by criticizing the opposed party rather than promoting the supporting party.

## 6.2 OSN Data Analysis

**Biased.** Wong et al. [191] collect and analyze 119M tweets pertaining to the 2012 US presidential election to quantify political leaning of users and news outlets. By formulating the problem as an ill-posed linear inverse problem, they propose an inference engine that considers tweeting behavior of articles. Having demonstrated their inference engine, the authors report results for the political leaning scores of news sources and users on Twitter. Golbeck and Hansen [74] provide a technique to estimate audience preferences in a given domain on Twitter, with a particular focus on political preferences. Different from methods that assess audience preference based on citation networks of news sources as a proxy, they directly measure the audience itself via their social network. Their technique is composed of three steps: 1) apply ground truth scores (they used Americans for Democratic Action reports as well as DW-Nominate scores) to a set of seed nodes in the network, 2) map these scores to the seed group’s followers to create “P-scores”, and 3) map the P-scores to the target of interest (e.g., government agencies or think tanks). One important take away from this work is that *Republicans are over-represented on Twitter with respect to their representation in Congress*, at least during the 2012 election cycle. To deal with this, they built a balanced dataset by randomly sampling from bins formed by the number of followers a seed group account had.

**Propaganda.** Jackson and Welles [92] demonstrate how Twitter can be exploited to organize and promote counter narratives. To



do so, they investigate the misuse of a Twitter hashtag (#myNYPD) during the 2014 New York City Police Department public relations campaign. In this campaign, this hashtag was greatly disseminated to promote counter narratives about racism and police misconduct. The authors leverage network and qualitative discourse analysis to study the structure and strategies used for promoting counterpublic narratives.

Hegelich and Janetzko [83] investigate whether bots on Twitter are used as political actors. By exposing and analyzing 1.7K bots on Twitter, during the Russian/Ukrainian conflict, they find that the botnet has a political agenda and that bots exhibit various behaviors. Specifically, they find that bots try to hide their identity, to be interesting by promoting topics through the use of hashtags and retweets. Howard and Kollanyi [90] focus on the 2016 UK referendum and the role of bots in the conversations on Twitter. They analyze 1.5M tweets from 313K Twitter accounts collected by searching specific hashtags related to the referendum. Their analysis indicates that most of the tweets are in favor of exiting the EU, there are bots with different levels of automation and that 1% of the accounts generate 33% of the overall messages. They also note that among the top sharers, there are a lot of bot accounts that are mostly retweeting and not generating new content. In a similar work, Howard et al. [89] study Twitter behavior during the second 2016 US Presidential Debate. They find that Twitter activity is more pro-Trump and that a lot of activity is driven by bots. However, they note that a substantial amount of tweets is original content posted from regular Twitter users. Woolley [192] analyzes several articles regarding the use of bots in OSNs for political purposes. Specifically, he undertakes a qualitative content analysis on 41 articles regarding political bots from various countries obtained from the Web. One of his main findings is that the use of bots varies from country to country and that some countries (e.g., Argentina, China, Russia, USA, etc.) use political bots on more than one type of event. For example, they report the use of Chinese political bots for elections, for protests and for security reasons.

In the Chinese political stage, during December 2014, an anonymous blogger released an archive of emails pertaining to the employment of Wumao, a group of people that gets paid to disseminate propaganda on social media, from the Chinese government. King et al. [103] analyzed these leaks and found out 43K posts that were posted by Wumao. Their main findings are: 1) by analyzing the time-series of these posts, they find bursty activity, hence signs of coordination of the posters; 2) most of the posters are individuals working for the government; and 3) by analyzing the content of the message, they note that posters usually post messages for distraction rather than discussions of controversial matters (i.e., supporting China's regime instead of discussing an event). Similarly to the previous work, Yang et al. [197] study the Wumao by analyzing 26M posts from 2.7M users on the Sina Weibo OSN, aiming to provide insights regarding the behavior and the size of Wumao. Due to the lack of ground truth data, they use clustering and topic modeling techniques, in order to cluster users that post politics-related messages with similar topics. By manually checking the users on the produced clusters, they conclude that users that post pro-government messages are distributed across multiple clusters, hence there is no signs of coordination of the Wumao on Sina Weibo for the period of their dataset (August 2012 and August 2013).

Zannettou et al. [201] study Russian state-sponsored troll accounts and measure the influence they had on Twitter and other Web communities. They find that Russian trolls were involved in the discussion of political events, and that they exhibit different behavior when compared to random users. Finally, they show that their influence was not substantial, with the exception of the dissemination of articles from state-sponsored Russian news outlets like Russia Today (RT). Allcot and Gentzkow [29] make a large scale analysis on Facebook during the period of the 2016 US election. Their results provide the following interesting statistics about the US election: 1) 115 pro-Trump fake stories are shared 30M times, whereas 41 pro-Clinton fake stories are shared 7.6M times. This indicates that fake news stories that favor Trump are more profound in Facebook. 2) The aforementioned 37.6M shares translates to 760M instances of a user clicking to the news articles. This indicates the high reachability of the fake news stories to end-users. 3) By undertaking a 1200-person survey, they highlight that a user's education, age and overall media consumption are the most important factors that determine whether a user can distinguish false headlines.

**Rumors.** Shin et al. [161] undertake a content-based analysis on 330K tweets pertaining to the 2012 US election. Their findings agree with existing literature, noting that users that spread rumors are mostly sharing messages against a political person. Furthermore, they highlight the resilience of rumors despite the fact that rumor debunking evidence was disseminated in Twitter; however, this does not apply for rumors that originate from satire websites.

### 6.3 Other models/algorithms

**Biased.** An et al. [31] study the interactions of 7M followers of 24 US news outlets on Twitter, in order to identify political leaning. To achieve this, they create a distance model, based on co-subscription relationships, that maps news sources to a dimensional dichotomous political spectrum. Also, they propose a real-time application, which utilizes the underlying model, and visualizes the ideology of the various news sources. Yang et al. [196] investigate the topics of discussions on Twitter for 51 US political persons, including President Obama. The main finding of this work is that Republicans and Democrats are similarly active on Twitter with the difference that Democrats tend to use hashtags more frequently. Furthermore, by utilizing a graph that demonstrates the similarity of the agenda of each political person, they highlight that Republicans are more clustered. This indicates that Republicans tend to share more tweets regarding their party's issues and agenda.

**Propaganda.** Al-khateeb and Agarwal [28] study the dissemination of propaganda on Twitter from terrorist organizations (

namely ISIS). They propose a framework based on social studies that aim to identify social and behavioral patterns of propaganda messages disseminated by a botnet. Their main findings are that bots exhibit similar behaviors (i.e., similar sharing patterns, similar usernames, lot of tweets in a short period of time) and that they share information that contains URLs to other sites and blogs. Ranganath et al. [149] focus on the detection of political advocates (individuals that use social media to strategically push a political agenda) on Twitter. The authors note that identifying advocates is not a straightforward task due to the nuanced and diverse message construction and propagation strategies. To overcome this, they propose a framework that aims to model all the different propagation and message construction strategies of advocates. Their evaluation on two datasets on Twitter regarding gun rights and elections demonstrate that the proposed framework achieves good performance with a 93% AUC score.

**Rumors.** Jin et al. [96] study the 2016 US Election through the Twitter activity of the followers of the two presidential candidates. For identifying rumors, they collect rumor articles from Snopes and then they use text similarity algorithms based on: 1) Term frequency-inverse document frequency (TF-IDF); 2) BM25 proposed in [153] 3) Word2Vec embeddings [127]; 4) Doc2Vec embeddings [113]; 5) Lexicon used in [205]. Their evaluation indicates that the best performance is achieved using the BM25-based approach. This algorithm is subsequently used to classify the tweets of the candidates' followers. Based on the predictions of the algorithm, their main findings are: 1) rumors are more prevalent during election period; 2) most of the rumors are posted by a small group of users; 3) rumors are mainly posted to debunk rumors that are against their presidential candidate, or to inflict damage on the other candidate; and 4) rumor sharing behavior increases in key points of the presidential campaign and in emergency events.

## 6.4 False information in political stage - Future Directions

The main insights from the review of work that focus on the political stage are: 1) Temporal analysis can be leveraged to assess coordination of bots, state-sponsored actors, and orchestrated efforts on disseminating political false information; 2) Bots are extensively used for the dissemination of political false information; 3) Machine learning techniques can assist in detecting political false information and political leaning of users. However, there are concerns about the generalization of such solutions on other datasets/domains; and 4) Political campaigns are responsible for the substantial dissemination of political false information in mainstream Web communities.

As future directions for understanding and mitigating the effects of false information on the Web, we propose the following. First, there is extensive anecdotal evidence highlighting that Web communities are used by state-sponsored troll factories, e.g., the recent news regarding Russian troll factories deployed to influence the outcomes of the 2016 Brexit [152] referendum and the 2016 US presidential election [159]. We thus propose investigating this phenomenon both from the perspective of user analytics as well as societal impact. Additionally, there is a lack of studies providing insight on how politics-related false information is disseminated across the Web; most studies focus on a single Web community or to specific events or do not examine politics. Understanding how political information propagates across the Web will help society identify the source of false information and lead to successful containment efforts.

## 7 Other related work

In this section we shall present work that is relevant to the false information ecosystem that does not fit in the aforementioned lines of work. Specifically, we group these studies in the following categories: 1) General Studies; 2) Systems; and 3) Use of images on the false information ecosystem.

### 7.1 General Studies

**Credibility Assessment.** Buntain and Golbeck [47] compare the accuracy of models that use features based on journalists assessments and crowdsourced assessments. They indicate that there is small overlap between the two features sets despite the fact that they provide statistically correlated results. This indicates that crowdsourcing workers discern different aspects of the stories when compared to journalists. Finally, they demonstrate that models that utilize features from crowdsourcing outperform the models that utilize features from journalists. Zhang et al. [203] present a set of indicators that can be used to assess the credibility of articles. To find these indicators they use a diverse set of experts (coming from multiple disciplines), which analyzed and annotated 40 news articles. Despite the low number of annotated articles, this inter-disciplinary study is important as it can help in defining standards for assessing the credibility of content on the Web. Mangolin et al. [122] study the interplay between fact-checkers and rumor spreaders on social networks finding that users are more likely to correct themselves if the correction comes from a user they follow when compared to a stranger.

**Conspiracy Theories.** Starbird [167] performs a qualitative analysis on Twitter regarding shooting events and conspiracy theories. Using graph analysis on the domains linked from the tweets, she provides insight on how various websites work to promote conspiracy theories and push political agendas.

**Fabricated.** Horne and Adah [88] focus on the headline of fake and real news. Their analysis on three datasets of news articles highlight that fake news have substantial differences in their structure when compared with real news. Specifically, they report that generally the structure of the content and the headline is different. That is, fake news are smaller in size, use simple words, and use longer and “clickbaity” headlines. Potts et al. [146] study Reddit and 4chan and how their interface is a part of their culture that affects their information sharing behavior. They analyze the information shared on these two platforms during the 2013 Boston Marathon bombings. Their findings highlight that users on both sites tried to find the perpetrator of the attack by creating conversations for the attack, usually containing false information. Bode and Vraga [40] propose a new function on Facebook, which allow users to observe related stories that either confirm or correct false information; they highlight that using this function users acquire a better understanding of the information and its credibility. Finally, Pennycook and Rand [139] highlight that by attaching warnings to news articles can help users to better assess the credibility of articles, however news articles that are not attached with warnings are considered as validated, which is not always true, hence users are tricked.

**Propaganda.** Chen et al. [54] study the behavior of hidden paid posters on OSNs. To better understand how these actors work, an author of this work posed as a hidden paid poster for a site [17] that gives users the option to be hidden paid posters. This task revealed valuable information regarding the organization of such sites and the behavior of the hidden paid posters, who are assigned with missions that need to be accomplished within a deadline. For example, a mission can be about posting articles of a particular content on different sites. A manager of the site can verify the completion of the task and then the hidden paid poster gets paid. To further study the problem, they collect data pertaining to a dispute between two big Chinese IT companies, from users of 2 popular Chinese news sites (namely Sohu [21] and Sina [18]). During this conflict there were strong suspicions that both companies employed hidden paid posters to disseminate false information that aimed to inflict damage to the other company. By undertaking statistical and semantic analysis on the hidden paid posters’ content they uncover a lot of useful features that can be used in identifying hidden paid posters. To this end, they propose the use of SVMs in order to detect such users by taking into consideration statistical and semantic features; their evaluation show that they can detect users with 88% accuracy.

**Rumors.** Starbird et al. [169] study and identify various types of expressed uncertainty within posts in OSN during a rumor’s lifetime. To analyze the uncertainty degree in messages, the paper acquires 15M tweets related to two crisis incidents (Boston Bombings and Sydney Siege). They find that specific linguistic patterns are used in rumor-related tweets. Their findings can be used in future detection systems in order to detect rumors effectively in a timely manner. Zubiaga et al. [210] propose a different approach in collecting and preparing datasets for false information detection. Instead of finding rumors from busting websites and then retrieving data from OSNs, they propose the retrieval of OSN data that will subsequently annotated by humans. In their evaluation, they retrieve tweets pertaining to the Ferguson unrest incident during 2014. They utilize journalists that act as annotators with the aim to label the tweets and their conversations. Specifically, the journalists annotated 1.1k tweets, which can be categorized into 42 different stories. Their findings show that 24.6% of the tweets are rumorous. Finally, Spiro et al. [166] undertake a quantitative analysis on tweets pertaining to the 2010 Deepwater Horizon oil spill. They note that media coverage increased the number of tweets related to the disaster. Furthermore, they observe that retweets are more commonly transmitted serially when they have event-related keywords.

## 7.2 Systems

**Biased.** Park et al. [137] note that biased information is profoundly disseminated in OSNs. To alleviate this problem, they propose NewsCube: a service that aims to provide end-users with all the different aspects of a particular story. In this way, end-users can read and understand the stories from multiple perspectives hence assisting in the formulation of their own unbiased view for the story. To achieve this, they perform structure-based extraction of the different aspects that exist in news stories. These aspects are then clustered in order to be presented to the end-users. To evaluate the effectiveness of their system, they undertake several user studies that aim to demonstrate the effectiveness in terms of the ability of the users to construct balanced views when using the platform. Their results indicate that 16 out of 33 participants stated that the platform helped them formulate a balanced view of the story, 2 out of 33 were negative, whereas the rest were neutral.

**Credibility Assessment.** Hassan et al. [81] propose FactWatcher, a system that reports facts that can be used as leads in stories. Their system is heavily based on a database and offers useful features to its users such as ranking of the facts, keyword-based search and fact-to-statement translation. Ennals et al. [64] describe the design and implementation of Dispute Finder, which is a browser extension that allows users to be warned about claims that are disputed by sources that they might trust. Dispute Finder maintains a database with well-known disputed claims which are used to inform end-users in real-time while they are reading stories. Users are also able to contribute to the whole process by explicitly flagging content as disputed, or as evidence to dispute other claims. In the case of providing evidence, the system requires a reference to a trusted source that supports the user’s actions, thus ensuring the quality of user’s manual annotations. Mitra and Gilbert [128] propose CREDBANK that aims to process large datasets by combining machine and human computations. The former is used to summarize tweets in events, while the latter is responsible for assessing the credibility of the content. Pirolli et al. [142] focus on Wikipedia and develop a system that presents users an interactive dashboard, which includes the history of article content and edits. The main finding is that users can better judge the credibility of an article, given that they are presented with the history of the article and edits through an

interactive dashboard.

### 7.3 Use of images on the false information ecosystem

Information can be disseminated via images on the Web. The use of images increases the credibility of the included information, as users tend to believe more information that is substantiated with an image. However, nowadays, images can be easily manipulated, hence used for the dissemination of false information. In this section, we provide an overview of the papers that studied the problem of false information on the Web, while considering images.

**Fabricated.** Boididou et al. [42, 41] focus on the use of multimedia in false information spread in OSNs. In [41] they prepare and propose a dataset of 12K tweets, which are manually labeled as fake, true, or unknown. A tweet is regarded as true if the image is referring to a particular event and fake if the image is not referring to a particular event. The authors argue that this dataset can help researchers in the task of automated identification of fake multimedia within tweets. In [42] they study the challenges that exist in providing an automated verification system for news that contain multimedia. To this end, they propose the use of conventional classifiers with the aim to discern fake multimedia pertaining to real events. Their findings demonstrate that generalizing is extremely hard as their classifiers perform poorly (58% accuracy) when they are trained with a particular event and they are tested with another. Diego Saez-Trumper [156] proposes a Web application, called Fake Tweet Buster, that aims to warn users about tweets that contain false information through images or users that habitually diffuse false information. The proposed approach is based on the reverse image search technique (using Google Images) in order to determine the origin of the image, its age and its context. Furthermore, the application considers user attributes and crowdsourcing data in order to find users that consistently share tweets that contain false information on images. Pasquini et al. [138] aim to provide image verification by proposing an empirical system that seeks visually and semantically related images on the web. Specifically, their system utilizes news articles metadata in order to search, using Google’s search engine, for relevant news articles. These images are then compared with the original’s article images in order to identify whether the images were tampered. To evaluate their approach, they created dummy articles with tampered images in order to simulate the whole procedure.

Jin et al. [100] emphasize the importance of images in news articles for distinguishing its truthfulness. They propose the use of two sets of features extracted from images in conjunction with features that are proposed by [51, 112]. For the image features, they define a set of visual characteristics as well as overall image statistics. Their data is based on a corpus obtained from the Sina Weibo that comprises 50K posts and 26K images. For evaluating the image feature set, they use conventional machine learning techniques: namely SVM, Logistic Regression, KStar, and Random Forest. They find that the proposed image features increase the accuracy by 7% with an overall accuracy of 83%. In a follow-up work, Jin et al. [98] leverage deep neural networks with the goal of distinguishing the credibility of images. They note that this task is extremely difficult as images can be misleading in many ways. Specifically, images might be outdated (i.e., old images that are falsely used to describe a new event), inaccurate, or even manipulated. To assess the image credibility, they train a Convolutional Neural Network (CNN) using a large-scale auxiliary dataset that comprises 600K labeled fake and real images. Their intuition is that the CNN can extract useful hyperparameters that can be used to detect eye-catching and visually striking images, which are usually used to describe false information. Their evaluation indicates that the proposed model can outperform several baselines in terms of the precision, recall, F1, and accuracy scores. Gupta et al. [78] focus on the diffusion of fake images in Twitter during Hurricane Sandy in 2012. They demonstrate that the use of automated techniques (i.e., Decision Trees) can assist in distinguishing fake images from real ones. Interestingly, they note that the 90% of the fake images came from the top 0.3% of the users.

## 8 Discussion & Conclusions

In this work, we have presented an overview of the false information ecosystem. Specifically, we have presented the various types of false information that can be found online, the different actors of the false information ecosystem as well as their motives for diffusing controversial information. Through the identification of several lines of work, we have presented the existing work on the false information ecosystem. Namely, we have presented studies on user perception, propagation dynamics, detection and containment of false information, as well as the dynamics of false information on the political stage. Also, we present some gaps of the existing literature that can be exploited by researchers in order to further study the increasing problem of false information on the Web.

To conclude, we share some thoughts about the problem of false information on the Web’s ecosystem. We argue that at the current stage current automated solutions, that do not use human input, are unable to effectively mitigate the problem of false information on the Web. Therefore, we feel that we should put extra effort in raising awareness of the problem to regular users of social networks, so that they can distinguish false information and potentially understand if a post is made from a legitimate user instead of a bot or state-sponsored actors. When it comes to scientific research and solutions for the problems, we argue that it is extremely important to tackle the problem from a holistic point of view. That is, researchers should take into account multiple Web communities when considering the problem of false information on the Web. Also, we should focus on providing models and methods that generalize well on other communities or datasets. Finally, researchers should focus on designing and

developing real-time platforms that will shed light about the propagation of false information across multiple Web communities. For instance, inform Twitter users that 4chan users are pushing a particular “hashtag” in the Twitter platform, with the goal of promoting information of questionable credibility.

## 9 Acknowledgments

This work is supported by the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie “ENCASE” project (Grant Agreement No. 691025).

## References

- [1] Amazon Mechanical Turk. <https://www.mturk.com/>.
- [2] CrowdFlower. <https://www.crowdfunder.com/>.
- [3] Definition of half-truth. <https://www.merriam-webster.com/dictionary/half-truth>.
- [4] Epinions. <http://www.epinions.com/>.
- [5] Extreme Value Theory. [https://en.wikipedia.org/wiki/Extreme\\_value\\_theory](https://en.wikipedia.org/wiki/Extreme_value_theory).
- [6] Facebook’s failure: did fake news and polarized politics get trump elected? <https://www.theguardian.com/technology/2016/nov/10/facebook-fake-news-election-conspiracy-theories>.
- [7] Fake news. its complicated. <https://firstdraftnews.com/fake-news-complicated/>.
- [8] False smartphone alert of huge earthquake triggers panic in japan. <https://www.theguardian.com/world/2016/aug/01/false-alert-of-huge-earthquake-triggers-panic-in-japan>.
- [9] French fear putin and trump followers are using 4chan to disrupt presidential election. <https://venturebeat.com/2017/05/05/french-fear-putin-and-trump-followers-are-using-4chan-to-disrupt-presidential-election/>.
- [10] H-index. <https://en.wikipedia.org/wiki/H-index>.
- [11] How isis and russia won friends and manufactured crowds. <https://www.wired.com/story/isis-russia-manufacture-crowds/>.
- [12] How russian trolls support of third parties could have cost hillary clinton the election. <https://qz.com/1210369/russia-donald-trump-2016-how-russian-trolls-support-of-us-third-parties-may-have-cost-hillary-clinton-the-election/>.
- [13] How trump consultants exploited the facebook data of millions. <https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html>.
- [14] The origins of writerly words. <http://time.com/82601/the-origins-of-writerly-words/>.
- [15] People shared nearly as much fake news as real news on twitter during the election. <https://qz.com/1090903/people-shared-nearly-as-much-fake-news-as-real-news-on-twitter-during-the-election/>.
- [16] SatireWire. <http://www.satirewire.com/>.
- [17] Shuijunwang. <http://www.shuijunwang.com>.
- [18] Sina. <http://www.sina.com/>.
- [19] SNAP Datasets. <http://snap.stanford.edu/data/>.
- [20] Snopes. <http://www.snopes.com/>.
- [21] Sohu. <http://www.sohu.com/>.
- [22] The Onion. <http://www.theonion.com/>.
- [23] Useful idiot wiki. [http://rationalwiki.org/wiki/Useful\\_Idiot](http://rationalwiki.org/wiki/Useful_Idiot).
- [24] “we actually elected a meme as president”: How 4chan celebrated trump’s victory”. ” <https://www.washingtonpost.com/news/the-intersect/wp/2016/11/09/we-actually-elected-a-meme-as-president-how-4chan-celebrated-trumps-victory/>”.
- [25] Xinhuanet. <http://www.xinhuanet.com/>.
- [26] Adperfect. How fake news is creating profits. <http://www.adperfect.com/how-fake-news-is-creating-profits/>, 2017.
- [27] S. Afroz, M. Brennan, and R. Greenstadt. Detecting hoaxes, frauds, and deception in writing style online. In *SP*, 2012.
- [28] S. Al-khateeb and N. Agarwal. Examining botnet behaviors for propaganda dissemination: A case study of isil’s beheading videos-based propaganda. In *ICDMW*, 2015.
- [29] H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. Technical report, National Bureau of Economic Research, 2017.
- [30] M. AlRubaian, M. Al-Qurishi, M. Al-Rakhami, S. M. M. Rahman, and A. Alamri. A multistage credibility analysis model for microblogs. In *ASONAM*, 2015.
- [31] J. An, M. Cha, K. P. Gummadi, J. Crowcroft, and D. Quercia. Visualizing media bias through Twitter. In *ICWSM*, 2012.
- [32] A. Anagnostopoulos, A. Bessi, G. Caldarelli, M. Del Vicario, F. Petroni, A. Scala, F. Zollo, and W. Quattrociocchi. Viral misinformation: The role of homophily and polarization. In *WWW Companion*, 2015.
- [33] A. Anand, T. Chakraborty, and N. Park. We used Neural Networks to Detect Clickbaits: You won’t believe what happened Next! In *ECIR*, 2017.

- [34] C. Andrews, E. Fichet, Y. Ding, E. S. Spiro, and K. Starbird. Keeping Up with the Tweet-Dashians: The Impact of Official Accounts on Online Rumoring. In *CSCW*, 2016.
- [35] A. Arif, K. Shanahan, F.-J. Chou, Y. Dosouto, K. Starbird, and E. S. Spiro. How information snowballs: Exploring the role of exposure in online rumor propagation. In *CSCW*, 2016.
- [36] A. Bessi. On the statistical properties of viral misinformation in online social media. In *Physica A*, 2016.
- [37] A. Bessi, M. Coletto, G. A. Davidescu, A. Scala, G. Caldarelli, and W. Quattrociocchi. Science vs conspiracy: Collective narratives in the age of misinformation. In *PLoS one*, 2015.
- [38] L. M. Bettencourt, A. Cintrón-Arias, D. I. Kaiser, and C. Castillo-Chavez. The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models. 2006.
- [39] P. Biyani, K. Tsioutsoulouklis, and J. Blackmer. “8 amazing secrets for getting more clicks”: Detecting clickbaits in news streams using article informality. 2016.
- [40] L. Bode and E. K. Vraga. In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication*, 2015.
- [41] C. Boididou, K. Andreadou, S. Papadopoulos, D.-T. Dang-Nguyen, G. Boato, M. Riegler, and Y. Kompatsiaris. Verifying multimedia use at mediaeval 2015. In *MediaEval*, 2015.
- [42] C. Boididou, S. Papadopoulos, Y. Kompatsiaris, S. Schifferes, and N. Newman. Challenges of computational verification in social multimedia. In *WWW*, 2014.
- [43] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu. The socialbot network: when bots socialize for fame and money. In *ACSAC*, 2011.
- [44] P. R. Brewer, D. G. Young, and M. Morreale. The impact of real news about “fake news”: Intertextual processes and political satire. In *IJPOR*, 2013.
- [45] C. Budak, D. Agrawal, and A. El Abbadi. Limiting the spread of misinformation in social networks. In *WWW*, 2011.
- [46] C. Budak, S. Goel, and J. M. Rao. Fair and balanced? quantifying media bias through crowdsourced content analysis. In *Public Opinion Quarterly*, 2016.
- [47] C. Buntain and J. Golbeck. I Want to Believe: Journalists and Crowdsourced Accuracy Assessments in Twitter. In *arXiv preprint arXiv:1705.01613*, 2017.
- [48] C. Burfoot and T. Baldwin. Automatic satire detection: Are you having a laugh? In *ACL-IJCNLP*, 2009.
- [49] J. K. Burgoon, J. Blair, T. Qin, and J. F. Nunamaker Jr. Detecting deception through linguistic analysis. In *ISI*, 2003.
- [50] W. J. Campbell. *Yellow journalism: Puncturing the myths, defining the legacies*. 2001.
- [51] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *WWW*, 2011.
- [52] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *ICWSM*, 2010.
- [53] A. Chakraborty, B. Paranjape, S. Kakarla, and N. Ganguly. Stop clickbait: Detecting and preventing clickbaits in online news media. In *ASONAM*, 2016.
- [54] C. Chen, K. Wu, V. Srinivasan, and X. Zhang. Battling the internet water army: Detection of hidden paid posters. In *ASONAM*, 2013.
- [55] X. Chen, S.-C. J. Sin, Y.-L. Theng, and C. S. Lee. Why Do Social Media Users Share Misinformation? In *JCDL*, 2015.
- [56] Y. Chen, N. J. Conroy, and V. L. Rubin. Misleading online content: Recognizing clickbait as false news. In *MDD*, 2015.
- [57] Y. Y. Chen, S.-P. Yong, and A. Ishak. Email hoax detection system using levenshtein distance method. In *JCP*, 2014.
- [58] M. Conover, J. Ratkiewicz, M. R. Francisco, B. Gonçalves, F. Menczer, and A. Flammini. Political Polarization on Twitter. In *ICWSM*, 2011.
- [59] M. Conti, D. Lain, R. Lazzeretti, G. Lovisotto, and W. Quattrociocchi. It’s Always April Fools’ Day! On the Difficulty of Social Network Misinformation Classification via Propagation Features. In *IEEE WIFS*, 2017.
- [60] A. Dang, A. Moh’d, E. Milios, and R. Minghim. What is in a rumour: Combined visual analysis of rumour flow and user activity. In *CGI*, 2016.
- [61] A. Dang, M. Smit, A. Moh’d, R. Minghim, and E. Milios. Toward understanding how users respond to rumours in social media. In *ASONAM*, 2016.
- [62] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi. The spreading of misinformation online. In *National Academy of Sciences*, 2016.
- [63] B. Doerr, M. Fouz, and T. Friedrich. Why Rumors Spread So Quickly in Social Networks. In *Commun. ACM*, 2012.
- [64] R. Ennals, B. Trushkowsky, and J. M. Agosta. Highlighting disputed claims on the web. In *WWW*, 2010.
- [65] European Union. General Data Protection Regulation. <https://gdpr-info.eu/>, 2018.
- [66] L. Fan, Z. Lu, W. Wu, B. Thuraisingham, H. Ma, and Y. Bi. Least cost rumor blocking in social networks. In *ICDCS*, 2013.
- [67] M. Farajtabar, J. Yang, X. Ye, H. Xu, R. Trivedi, E. Khalil, S. Li, L. Song, and H. Zha. Fake News Mitigation via Point Process Based Intervention. In *JMLR*, 2017.
- [68] L. Feldman. Partisan differences in opinionated news perceptions: A test of the hostile media effect. In *Political Behavior*, 2011.
- [69] M. Fenster. *Conspiracy theories: Secrecy and power in American culture*. U of Minnesota Press, 1999.
- [70] E. Ferrara, O. Varol, F. Menczer, and A. Flammini. Detection of promoted social media campaigns. In *ICWSM*, 2016.

- [71] S. Finn, P. T. Metaxas, and E. Mustafaraj. Investigating rumor propagation with twittertrails. In *Computation and Journalism*, 2014.
- [72] A. Friggeri, L. A. Adamic, D. Eckles, and J. Cheng. Rumor cascades. In *ICWSM*, 2014.
- [73] G. Giasemidis, C. Singleton, I. Agrafiotis, J. R. Nurse, A. Pilgrim, C. Willis, and D. V. Greetham. Determining the veracity of rumours on twitter. In *SocInfo*, 2016.
- [74] J. Golbeck and D. L. Hansen. A method for computing political preference among twitter followers. In *Social Networks*, 2014.
- [75] A. Gupta and P. Kumaraguru. Credibility ranking of tweets during high impact events. In *PSOSM*, 2012.
- [76] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier. Tweetcred: Real-time credibility assessment of content on twitter. In *SocInfo*, 2014.
- [77] A. Gupta, H. Lamba, and P. Kumaraguru. \$1.00 per rt# bostonmarathon# prayforboston: Analyzing fake content on twitter.
- [78] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *WWW*, 2013.
- [79] S. Hamidian and M. T. Diab. Rumor identification and belief investigation on twitter. In *NAACL-HLT*, 2016.
- [80] J. T. Hancock, L. E. Curry, S. Goorha, and M. Woodworth. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. In *Discourse Processes*, 2007.
- [81] N. Hassan, A. Sultana, Y. Wu, G. Zhang, C. Li, J. Yang, and C. Yu. Data in, fact out: automated monitoring of facts by factwatcher. In *VLDB Endowment*, 2014.
- [82] Z. He, Z. Cai, and X. Wang. Modeling propagation dynamics and developing optimized countermeasures for rumor spreading in online social networks. In *ICDCS*, 2015.
- [83] S. Hegelich and D. Janetzko. Are social bots on twitter political actors? empirical evidence from a ukrainian social botnet. In *ICWSM*, 2016.
- [84] S. Heller. Bat Boy, Hillary Clinton's Alien Baby, and a Tabloid's Glorious Legacy. <https://www.theatlantic.com/entertainment/archive/2014/10/the-ingenious-sensationalism-of-the-weekly-world-new/381525/>, 2014.
- [85] E. Higgins. Fake news is spiraling out of control - and it is up to all of us to stop it. <https://www.ibtimes.co.uk/fake-news-spiralling-out-control-it-all-us-stop-it-1596911>, 2016.
- [86] G. E. Hine, J. Onalapo, E. D. Cristofaro, N. Kourtellis, I. Leontiadis, R. Samaras, G. Stringhini, and J. Blackburn. Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. In *ICWSM*, 2017.
- [87] T. Hogg and K. Lerman. Social dynamics of digg. In *EPJ Data Science*, 2012.
- [88] B. D. Horne and S. Adali. This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News. In *ICWSM Workshop*, 2017.
- [89] P. Howard, B. Kollanyi, and S. Woolley. Bots and automation over twitter during the second us presidential debate. 2016.
- [90] P. N. Howard and B. Kollanyi. Bots, #strongerin, and #brexit: computational propaganda during the uk-eu referendum. In *arXiv preprint arXiv:1606.06356*, 2016.
- [91] Y. L. Huang, K. Starbird, M. Orand, S. A. Stanek, and H. T. Pedersen. Connected through crisis: Emotional proximity and the spread of misinformation online. In *CSCW*, 2015.
- [92] S. J. Jackson and B. Foucault Welles. Hijacking#mynypd: Social media dissent and networked counterpublics. In *Journal of Communication*, 2015.
- [93] E. Jaho, E. Tzoannos, A. Papadopoulos, and N. Sarris. Alethiometer: a framework for assessing trustworthiness and content validity in social media. In *WWW*, 2014.
- [94] F. Jin, E. Dougherty, P. Saraf, Y. Cao, and N. Ramakrishnan. Epidemiological modeling of news and rumors on twitter. In *SNA-KDD*, 2013.
- [95] F. Jin, W. Wang, L. Zhao, E. Dougherty, Y. Cao, C.-T. Lu, and N. Ramakrishnan. Misinformation propagation in the age of twitter. In *Computer*, 2014.
- [96] Z. Jin, J. Cao, H. Guo, Y. Zhang, Y. Wang, and J. Luo. Rumor detection on twitter pertaining to the 2016 us presidential election. In *arXiv preprint arXiv:1701.06250*, 2017.
- [97] Z. Jin, J. Cao, Y.-G. Jiang, and Y. Zhang. News credibility evaluation on microblog with a hierarchical propagation model. In *ICDM*, 2014.
- [98] Z. Jin, J. Cao, J. Luo, and Y. Zhang. Image Credibility Analysis with Effective Domain Transferred Deep Networks. In *arXiv preprint arXiv:1611.05328*, 2016.
- [99] Z. Jin, J. Cao, Y. Zhang, and J. Luo. News verification by exploiting conflicting social viewpoints in microblogs. In *AAAI*, 2016.
- [100] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian. Novel visual and statistical image features for microblogs news verification. In *ToM*, 2016.
- [101] G. S. Jowett and V. O'donnell. *Propaganda & persuasion*. Sage, 2014.
- [102] J.-H. Kim and G.-W. Bock. A study on the factors affecting the behavior of spreading online rumors: Focusing on the rumor recipient's emotions. In *PACIS*, 2011.
- [103] G. King, J. Pan, and M. E. Roberts. How the chinese government fabricates social media posts for strategic distraction, not engaged argument. In *Harvard University*, 2016.
- [104] R. E. Kopp. Pontryagin maximum principle. In *Mathematics in Science and Engineering*, 1962.
- [105] M. Koppel, J. Schler, and E. Bonchek-Dokow. Measuring differentiability: Unmasking pseudonymous authors. In *Machine Learning*

Research, 2007.

- [106] B. Kotnis and J. Kuri. Cost effective rumor containment in social networks. In *arXiv preprint arXiv:1403.6315*, 2014.
- [107] K. K. Kumar and G. Geethakumari. Detecting misinformation in online social networks using cognitive psychology. In *HCIS*, 2014.
- [108] S. Kumar and N. Shah. False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559*, 2018.
- [109] S. Kumar, R. West, and J. Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *WWW*, 2016.
- [110] S. Kwon, M. Cha, and K. Jung. Rumor detection over varying time windows. In *PLOS ONE*, 2017.
- [111] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang. Aspects of rumor spreading on a microblog network. In *SocInfo*, 2013.
- [112] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang. Prominent features of rumor propagation in online social media. In *ICDM*, 2013.
- [113] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *ICML*, 2014.
- [114] S. T. Lee. Lying to tell the truth: Journalists and the social context of deception. In *Mass Communication & Society*, 2004.
- [115] G. Liang, W. He, C. Xu, L. Chen, and J. Zeng. Rumor Identification in Microblogging Systems Based on Users' Behavior. In *IEEE Transactions on Computational Social Systems*, 2015.
- [116] X. Liu, A. Nourbakhsh, Q. Li, R. Fang, and S. Shah. Real-time Rumor Debunking on Twitter. In *CIKM*, 2015.
- [117] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha. Detecting rumors from microblogs with recurrent neural networks. In *IJCAI*, 2016.
- [118] J. Ma, W. Gao, Z. Wei, Y. Lu, and K.-F. Wong. Detect rumors using time series of social context information on microblogging websites. In *CIKM*, 2015.
- [119] J. Ma and D. Li. Rumor Spreading in Online-Offline Social Networks. 2016.
- [120] C. Maigrot, V. Claveau, E. Kijak, and R. Sicre. Mediaeval 2016: A multimodal system for the verifying multimedia use task. In *MediaEval*, 2016.
- [121] R. Marchi. With facebook, blogs, and fake news, teens reject journalistic "objectivity". 2012.
- [122] D. B. Margolin, A. Hannak, and I. Weber. Political fact-checking on twitter: when do corrections have an effect? *Political Communication*, 2018.
- [123] R. McCreadie, C. Macdonald, and I. Ounis. Crowdsourced rumour identification during emergencies. In *WWW*, 2015.
- [124] Medium. Different Examples of Propaganda in Social Media. <https://medium.com/@VasquezNnenna/different-examples-of-propaganda-in-social-media> 2018.
- [125] M. Mendoza, B. Poblete, and C. Castillo. Twitter under crisis: Can we trust what we rt? In *SOMA-KDD*, 2010.
- [126] T. Mihaylov, G. Georgiev, and P. Nakov. Finding Opinion Manipulation Trolls in News Community Forums. In *CoNLL*, 2015.
- [127] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 2013.
- [128] T. Mitra and E. Gilbert. CREDBANK: A Large-Scale Social Media Corpus With Associated Credibility Annotations. In *ICWSM*, 2015.
- [129] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. In *Random structures & algorithms*, 1995.
- [130] M. R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz. Tweeting is believing?: understanding microblog credibility perceptions. In *CSCW*, 2012.
- [131] A. Nadamoto, M. Miyabe, and E. Aramaki. Analysis of microblog rumors and correction texts for disaster situations. In *iiWAS*, 2013.
- [132] K. Napley. The Impact of Fake News: Politics. <https://www.lexology.com/library/detail.aspx?g=6c63091c-e81f-4512-8c47-52leadce65ff>, 2017.
- [133] D. T. Nguyen, N. P. Nguyen, and M. T. Thai. Sources of misinformation in online social networks: Who to suspect? In *MILCOM*, 2012.
- [134] N. P. Nguyen, G. Yan, M. T. Thai, and S. Eidenbenz. Containment of misinformation spread in online social networks. In *WebSci*, 2012.
- [135] O. Oh, K. H. Kwon, and H. R. Rao. An exploration of social media in extreme events: Rumor theory and twitter during the haiti earthquake. In *ICIS*, 2010.
- [136] P. Ozturk, H. Li, and Y. Sakamoto. Combating rumor spread on social media: The effectiveness of refutation and warning. In *HICSS*, 2015.
- [137] S. Park, S. Kang, S. Chung, and J. Song. Newscube: delivering multiple aspects of news to mitigate media bias. In *CHI*, 2009.
- [138] C. Pasquini, C. Brunetta, A. F. Vinci, V. Conotter, and G. Boato. Towards the verification of image integrity in online news. In *ICMEW*, 2015.
- [139] G. Pennycook and D. G. Rand. The implied truth effect: Attaching warnings to a subset of fake news stories increases perceived accuracy of stories without warnings. 2017.
- [140] W. A. Peterson and N. P. Gist. Rumor and public opinion. In *American Journal of Sociology*, 1951.
- [141] Y. Ping, Z. Cao, and H. Zhu. Sybil-aware least cost rumor blocking in social networks. In *GLOBECOM*, 2014.
- [142] P. Pirolli, E. Wollny, and B. Suh. So you know you're getting the best possible information: a tool that increases wikipedia credibility. In *CHI*. ACM.
- [143] Politifact. The more outrageous, the better: How clickbait ads make money for fake news sites. <http://www.politifact.com/punditfact/article/2017/oct/04/more-outrageous-better-how-clickbait-ads-make-mone/>, 2017.
- [144] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein. A Stylometric Inquiry into Hyperpartisan and Fake News. In *arXiv*



preprint *arXiv:1702.05638*, 2017.

- [145] M. Potthast, S. Köpsel, B. Stein, and M. Hagen. Clickbait detection. In *ECIR*, 2016.
- [146] L. Potts and A. Harrison. Interfaces as rhetorical constructions: reddit and 4chan during the boston marathon bombings. In *SIGDOC*, 2013.
- [147] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei. Rumor has it: Identifying misinformation in microblogs. In *EMNLP*, 2011.
- [148] Y. Qin, D. Wurzer, V. Lavrenko, and C. Tang. Spotting Rumors via Novelty Detection. In *arXiv preprint arXiv:1611.06322*, 2016.
- [149] S. Ranganath, X. Hu, J. Tang, and H. Liu. Understanding and identifying advocates for political campaigns on social media. In *WSDM*, 2016.
- [150] J. Ratkiewicz, M. Conover, M. R. Meiss, B. Gonçalves, A. Flammini, and F. Menczer. Detecting and tracking political abuse in social media. In *ICWSM*, 2011.
- [151] P. Resnick, S. Carton, S. Park, Y. Shen, and N. Zeffer. Rumorlens: A system for analyzing the impact of rumors and corrections in social media. In *Proc. Computational Journalism Conference*, 2014.
- [152] A. H. Robert Booth, Matthew Weaver and S. Walker. Russia used hundreds of fake accounts to tweet about brexit, data shows. <https://www.theguardian.com/world/2017/nov/14/how-400-russia-run-fake-accounts-posted-bogus-brexit-tweets>, 2017.
- [153] S. Robertson, H. Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. In *Foundations and Trends® in Information Retrieval*, 2009.
- [154] V. L. Rubin, N. J. Conroy, and Y. Chen. Towards news verification: Deception detection methods for news discourse. 2015.
- [155] V. L. Rubin, N. J. Conroy, Y. Chen, and S. Cornwell. Fake news or truth? using satirical cues to detect potentially misleading news. In *NAACL-HLT*, 2016.
- [156] D. Saez-Trumper. Fake tweet buster: a webtool to identify users promoting fake news on twitter. In *HT*, 2014.
- [157] E. Seo, P. Mohapatra, and T. Abdelzaher. Identifying rumors and their sources in social networks. In *SPIE defense, security, and sensing*, 2012.
- [158] D. Shah and T. Zaman. Rumors in a network: Who’s the culprit? In *IEEE Transactions on information theory*, 2011.
- [159] S. Shane. The fake americans russia created to influence the election. <https://www.nytimes.com/2017/09/07/us/politics/russia-facebook-twitter-election> 2017.
- [160] C. Shao, G. L. Ciampaglia, A. Flammini, and F. Menczer. Hoaxy: A platform for tracking online misinformation. In *WWW*, 2016.
- [161] J. Shin, L. Jian, K. Driscoll, and F. Bar. Political rumoring on twitter during the 2012 us presidential election: Rumor diffusion and correction. In *new media & society*, 2016.
- [162] H. Situngkir. Spread of hoax in social media. In *MPRA Paper 30674*, 2011.
- [163] Snopes. Boston Marathon Bombing Rumors. <https://www.snopes.com/fact-check/boston-marathon-bombing-rumors/>, 2013.
- [164] Snopes. Adam Sandler Death Hoax. <https://www.snopes.com/fact-check/adam-sandler-death-hoax-2/>, 2017.
- [165] P. Snyder, P. Doerfler, C. Kanich, and D. McCoy. Fifteen minutes of unwanted fame: Detecting and characterizing doxing. In *IMC*, 2017.
- [166] E. S. Spiro, S. Fitzhugh, J. Sutton, N. Pierski, M. Greczek, and C. T. Butts. Rumoring during extreme events: A case study of deepwater horizon 2010. In *WebSci*, 2012.
- [167] K. Starbird. Examining the Alternative Media Ecosystem through the Production of Alternative Narratives of Mass Shooting Events on Twitter. In *ICWSM*, 2017.
- [168] K. Starbird, J. Maddock, M. Orand, P. Achterman, and R. M. Mason. Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing. In *iConference*, 2014.
- [169] K. Starbird, E. Spiro, I. Edwards, K. Zhou, J. Maddock, and S. Narasimhan. Could this be true?: I think so! expressed uncertainty in online rumoring. In *CHI*, 2016.
- [170] E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret, and L. de Alfaro. Some Like it Hoax: Automated Fake News Detection in Social Networks. In *SoGood Workshop*, 2017.
- [171] M. Tambuscio, G. Ruffo, A. Flammini, and F. Menczer. Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks. In *WWW*, 2015.
- [172] The Independent. St Petersburg ‘troll farm’ had 90 dedicated staff working to influence US election campaign. <https://ind.pn/2yuCQdy>, 2017.
- [173] The New Yorker. How the NRA Manipulates Gun Owners and the Media. <https://www.newyorker.com/news/news-desk/how-the-nra-manipulates-gun-2017>.
- [174] R. Thomson, N. Ito, H. Suda, F. Lin, Y. Liu, R. Hayasaka, R. Isochi, and Z. Wang. Trusting tweets: The fukushima disaster and information source credibility on twitter. In *ISCRAM*, 2012.
- [175] G. Timberg. Spreading fake news becomes standard practice for governments across the world. <https://www.washingtonpost.com/news/the-switch/wp/2017/07/17/spreading-fake-news-becomes-standard-practice-for-governments-across-the-world-2017>.
- [176] G. Tong, W. Wu, L. Guo, D. Li, C. Liu, B. Liu, and D.-Z. Du. An Efficient Randomized Algorithm for Rumor Blocking in Online Social Networks. In *INFOCOM*, 2017.
- [177] R. M. Tripathy, A. Bagchi, and S. Mehta. A study of rumor control strategies on social networks. In *CIKM*, 2010.

- [178] US House of Representatives. Exposing Russias Effort to Sow Discord Online: The Internet Research Agency and Advertisements. <https://democrats-intelligence.house.gov/social-media-content/>, 2018.
- [179] S. Volkova, K. Shaffer, J. Y. Jang, and N. Hodas. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter.
- [180] S. Vosoughi and D. Roy. A human-machine collaborative system for identifying rumors on twitter. In *ICDMW*, 2015.
- [181] S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 2018.
- [182] M. Vuković, K. Pripužić, and H. Belani. An intelligent automatic hoax detection system. In *KES*, 2009.
- [183] B. Wang, G. Chen, L. Fu, L. Song, X. Wang, and X. Liu. Drimux: Dynamic rumor influence minimization with user experience in social networks. In *AAAI*, 2016.
- [184] N. Wang, L. Yu, N. Ding, and D. Yang. Containment of Misinformation Propagation in Online Social Networks with given Deadline. In *PACIS*, 2014.
- [185] W. Y. Wang. “Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection. In *arXiv preprint arXiv:1705.00648*, 2017.
- [186] Z. Wang, W. Dong, W. Zhang, and C. W. Tan. Rumor source detection with multiple observations: Fundamental limits and algorithms. In *ACM PER*, 2014.
- [187] Wikipedia. Murder of Seth Rich. [https://en.wikipedia.org/wiki/Murder\\_of\\_Seth\\_Rich](https://en.wikipedia.org/wiki/Murder_of_Seth_Rich), 2017.
- [188] Wikipedia. Pizzagate conspiracy theory. [https://en.wikipedia.org/wiki/Pizzagate\\_conspiracy\\_theory](https://en.wikipedia.org/wiki/Pizzagate_conspiracy_theory), 2017.
- [189] Wikipedia. Sandy Hook Elementary School shooting conspiracy theories. [https://en.wikipedia.org/wiki/Sandy\\_Hook\\_Elementary\\_School\\_shooting-conspiracy\\_theories](https://en.wikipedia.org/wiki/Sandy_Hook_Elementary_School_shooting-conspiracy_theories), 2018.
- [190] S. Wineburg and S. McGrew. Lateral reading: Reading less and learning more when evaluating digital information. 2017.
- [191] F. M. F. Wong, C. W. Tan, S. Sen, and M. Chiang. Quantifying Political Leaning from Tweets and Retweets. In *ICWSM*, 2013.
- [192] S. C. Woolley. Automating power: Social bot interference in global politics. In *First Monday*, 2016.
- [193] K. Wu, S. Yang, and K. Q. Zhu. False rumors detection on sina weibo by propagation structures. In *ICDE*, 2015.
- [194] L. Wu, J. Li, X. Hu, and H. Liu. Gleaning wisdom from the past: Early detection of emerging rumors in social media. In *SDM*, 2017.
- [195] F. Yang, Y. Liu, X. Yu, and M. Yang. Automatic detection of rumor on sina weibo. In *KDD*, 2012.
- [196] X. Yang, B.-C. Chen, M. Maity, and E. Ferrara. Social politics: Agenda setting and political communication on social media. In *SocInfo*, 2016.
- [197] X. Yang, Q. Yang, and C. Wilson. Penny for your thoughts: Searching for the 50 cent party on sina weibo. In *ICWSM*, 2015.
- [198] Z. Yang, C. Wang, F. Zhang, Y. Zhang, and H. Zhang. Emerging rumor identification for social media with hot topic detection. In *WISA*, 2015.
- [199] S. Zannettou, B. Bradlyn, E. De Cristofaro, M. Sirivianos, G. Stringhini, H. Kwak, and J. Blackburn. What is gab? a bastion of free speech or an alt-right echo chamber? In *WWW Companion*, 2018.
- [200] S. Zannettou, T. Caulfield, E. D. Cristofaro, N. Kourtellis, I. Leontiadis, M. Sirivianos, G. Stringhini, and J. Blackburn. The web centipede: Understanding how web communities influence each other through the lens of mainstream and alternative news sources. In *IMC*, 2017.
- [201] S. Zannettou, T. Caulfield, E. De Cristofaro, M. Sirivianos, G. Stringhini, and J. Blackburn. Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web. *arXiv preprint arXiv:1801.09288*, 2018.
- [202] S. Zannettou, S. Chatzis, K. Papadamou, and M. Sirivianos. The Good, the Bad and the Bait: Detecting and Characterizing Clickbait on YouTube. In *IEEE DLS*, 2018.
- [203] A. X. Zhang, A. Ranganathan, S. E. Metz, S. Appling, C. M. Sehat, N. Gilmore, N. B. Adams, E. Vincent, J. Lee, M. Robbins, et al. A structured response to misinformation: defining and annotating credibility indicators in news articles. In *WWW Companion*, 2018.
- [204] Q. Zhang, S. Zhang, J. Dong, J. Xiong, and X. Cheng. Automatic detection of rumor on social network. In *NLPCC*, 2015.
- [205] Z. Zhao, P. Resnick, and Q. Mei. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *WWW*, 2015.
- [206] D. X. Zhou, P. Resnick, and Q. Mei. Classifying the Political Leaning of News Articles and Users from User Votes. In *ICWSM*, 2011.
- [207] X. Zhou, J. Cao, Z. Jin, F. Xie, Y. Su, D. Chu, X. Cao, and J. Zhang. Real-time news certification system on sina weibo. In *WWW*, 2015.
- [208] F. Zollo, A. Bessi, M. Del Vicario, A. Scala, G. Caldarelli, L. Shekhtman, S. Havlin, and W. Quattrociocchi. Debunking in a World of Tribes. In *PloS one*, 2017.
- [209] F. Zollo, P. K. Novak, M. Del Vicario, A. Bessi, I. Mozetič, A. Scala, G. Caldarelli, and W. Quattrociocchi. Emotional dynamics in the age of misinformation. In *PloS one*, 2015.
- [210] A. Zubiaga, M. Liakata, R. Procter, K. Bontcheva, and P. Tolmie. Towards detecting rumours in social media. In *AAAIW*, 2015.
- [211] A. Zubiaga, M. Liakata, R. Procter, G. W. S. Hoi, and P. Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. In *PloS one*, 2016.