MIS2: Misinformation and Misbehavior Mining on the Web

5 authors, including:

Some of the authors of this publication are also working on these related projects:

Social Computing View project

GNNExplainer: Generating Explanations for Graph Neural Networks View project

False Information on Web and Social Media: A Survey

SRIJAN KUMAR, Computer Science, Stanford University, USA

NEIL SHAH*, Computer Science, Carnegie Mellon University, USA

False information can be created and spread easily through the web and social media platforms, resulting in widespread real-world impact. Characterizing how false information proliferates on social platforms and why it succeeds in deceiving readers are critical to develop efficient detection algorithms and tools for early detection. A recent surge of research in this area has aimed to address the key issues using methods based on feature engineering, graph mining, and information modeling. Majority of the research has primarily focused on two broad categories of false information: opinion-based (e.g., fake reviews), and fact-based (e.g., false news and hoaxes). Therefore, in this work, we present a comprehensive survey spanning diverse aspects of false information, namely (i) the actors involved in spreading false information, (ii) rationale behind successfully deceiving readers, (iii) quantifying the impact of false information, (iv) measuring its characteristics across different dimensions, and finally, (iv) algorithms developed to detect false information. In doing so, we create a unified framework to describe these recent methods and highlight a number of important directions for future research.[1]

Additional Key Words and Phrases: misinformation, fake news, fake reviews, rumors, hoaxes, web, internet, social media, social networks, bad actors, bots, propaganda, conspiracy, knowledge bases, e-commerce, disinformation, impact, mechanism, rationale, detection, prediction

1 INTRODUCTION

The web provides a highly interconnected world-wide platform for everyone to spread information to millions of people in a matter of few minutes, at little to no cost [12]. While it has led to ground-breaking phenomenon such as real-time citizen journalism [34], at the same time it has led to increased visibility and impact of both true and false information [57]. False information on the web and social media has affected stock markets [17], slowed responses during disasters [32], and terrorist attacks [27, 96]. Recent surveys have alarmingly shown that people increasingly get their news from social media than from traditional news sources [75, 88], making it of paramount importance to curtail false information on such platforms. With primary motives of influencing opinions and earning money [1, 46, 56, 94], the wide impact of false information makes it one of the modern dangers to society, according to the World Economic Forum [39]. Understanding the reasons for why and how false information is created is important to proactively detect it and mitigate its impact. In this survey, we review the state of the art scientific literature on false information on the web and social media to give a comprehensive description of its **mechanisms, rationale, impact, characteristics**, and **detection**. While recent surveys have focused on fake

---

*Dr. Shah is now at Snap Inc.

[1]A previous version of this survey will appear in the book titled Social Media Analytics: Advances and Applications, by CRC press, in 2018.

Authors' addresses: Srijan Kumar, Computer Science, Stanford University, USA, srijan@cs.stanford.edu; Neil Shah, Computer Science, Carnegie Mellon University, USA, neil@cs.cmu.edu.

---

news in social media [90, 120], the current survey broadly focuses on three types of false information on the web and social media—**fake reviews in e-commerce platforms, hoaxes on collaborative platforms, and fake news in social media.**

For ease of explanation and understanding, we categorize false information based on its *intent* and *knowledge* content. We also broadly focus on false information that is public and targets many people at the same time, such as false reviews or fake tweets, as opposed to targeted false information as in cases of scam. According to intent, false information can be categorized as *misinformation*, which is created *without* the intent to mislead, and *disinformation*, which is created *with* the intent of misleading and deceiving the reader [25, 35]. Both have negative influences, but the latter is arguably more dangerous as its creator's primary aim is expressly malicious. Based on knowledge, false information is categorized as ***opinion-based***, where a unique ground truth does not exist as in cases of reviewing products on e-commerce websites, or as ***fact-based***, which consists of lies about entities that have unique ground truth value [101]. We study both these types of false information in this survey.

The ***bad actors involved*** in creating and spreading false information on a large scale use armies of fake accounts, such as bots and sockpuppets [26, 47, 85, 97]. These accounts are synthetically created or compromised [85], controlled by a single underlying entity, and engineer the spread of false information through social networks [11, 19], with the aim of creating an illusion of public consensus towards the false pieces of information. Bots operate on a large scale for two purposes: first, to spread the same content, e.g., by retweeting, to a large audience, and second, by following each other to increase the social status of the accounts and apparent trustworthiness of information [26, 97]. Sockpuppet accounts engage with ordinary users in online discussions and agree with each other to amplify their point of view and oppose those who disagree with the information [47]. Alarmingly, bots and sockpuppets hold central locations in information networks and therefore, are in key positions to spread false information. We dig deep into the mechanisms of false information and the actors involved in Section 3.1.

False information would be ineffective if readers were able to easily identify that it is false and just discard it. However, the ***rationale behind successful deception*** by false information is evident from several research studies which show that humans are actually poor judges of false information [50, 70, 74, 114]. Specifically, humans are able to identify false information with accuracies between 53% and 78% across experiments with different types of false behaviors, including hoaxes, fake reviews, and fake news. Both trained and casual readers get fooled into believing false information when it is well written, long, and is well-referenced. Moreover, technological effects such as content personalization can lead to the creation of ideological echo chambers, so that people would receive the same false information multiple times through different channels and could even make it "go viral". Biases in information consumers (e.g., confirmation bias), lack of education, and low media consumption lead to people being deceived by false information. Further details of the rationale of deception using false information are explained in Section 3.2.

The spread of false information can have ***far-reaching impact***. Several research studies have measured the impact of false information in social media in terms of user engagement metrics, such as the number of likes, reshares, and pre-removal lifetime, for hoaxes [50], fake news [32, 92], and rumors [30]. They found that a small fraction of false information stories is highly impactful—they are liked, shared, and commented on more, generate deeper cascades of reshares than true information pieces, survive for a long time, and spread across the web. This high engagement of false information with readers shows the degree of impact it can have on public opinion and ideological perception. We discuss the impact of false information in detail in Section 4.

In response, considerable research has been conducted to both investigate and use the ***characteristics of false information*** to predict the veracity of new content. Fake reviews [43, 53, 61, 70, 82, 84], hoaxes [50], and fake news [13, 30, 37, 74, 92, 121] have been characterized using their textual content, temporal features, ratings, references, user properties, network properties, spreading behavior, and mitigation behavior. Along these characteristics, false information differs significantly from real information. For instance, the text is generally

Social Platform	Research papers
Twitter	Bessi et al. [13], Ferrara et al. [26], Gupta et al. [32], Howard et al. [38], Jin et al. [41, 42], Kim et al. [44], Mendoza et al. [57], Mitra et al. [59, 60], Nied et al. [66], Qazvinian et al. [77], Ruchansky et al. [81], Shah et al. [85], Shao et al. [86, 87], Starbird et al. [95, 96], Subrahmanian et al. [97], Tripathy et al. [103], Vosoughi et al. [105], Zeng et al. [118], Zubiaga et al. [121]
Facebook	Beutel et al. [16], Del et al. [22], Friggeri et al. [30], Nguyen et al. [64], Silverman et al. [91, 92], Tacchini et al. [100]
Review platforms	Akoglu et al. [5], Beutel et al. [15], Harris et al. [33], Hooi et al. [36], Jindal et al. [43], Kumar et al. [48], Li et al. [51–54], Lin et al. [55], Luca et al. [56], Minnich et al. [58], Mukherjee et al. [61, 62], Ott et al. [69, 70], Rayana et al. [79], Sandulescu et al. [82], Shah et al. [84], Wang et al. [106], Xie et al. [111], Yao et al. [114], Ye et al. [115]
Sina Weibo	Jiang et al. [40], Kim et al. [44], Ruchansky et al. [81], Wu et al. [110], Yang et al. [112]
Multi-platform	Reddit+Twitter+4chan: Zannettou et al. [117]
Other	Fake news articles: Horne et al. [37], Silverman et al. [92], Rubin et al. [80], Perez et al. [74], Wikipedia: Kumar et al. [50], False information websites: Albright et al. [7, 8], Fact checking website: Shu et al. [89] and Wang et al. [107], Crowdsourcing and tabloid websites: Perez et al. [74]

Table 1. This table categorizes research on false information based on the platforms they study.

longer, more exaggerated, and more opinionated compared to real reviews. Temporally, fake reviews are created in short bursts, i.e., several fake reviews are usually written by the same account or group of accounts in a short time period. The users who write these fake reviews and hoaxes are typically relatively new accounts with fewer reviews, and their local networks are often highly dense or overlapping. Additionally, majority of fake news is spread by a very small number of users and it spreads rapidly during its initial release, before it is even debunked. The characteristics of false information are discussed extensively in Section 5.

Finally, several algorithms have been created for effective ***detection of false information*** from its true counterparts. These algorithms can broadly be categorized into three categories: feature-based, graph-based, and propagation-modeling based. Feature-based algorithms leverage the unique characteristics for detection by using them as features in a machine learning model or rule-based framework [32, 37, 43, 50, 70, 77, 82]. Graph-based algorithms are developed to identify dense-blocks or dense subgraphs of users and information in the network [6, 16, 40, 48, 79, 106]. Propagation-modeling algorithms create information spread models for true information and use these models to identify false information [3, 18, 41, 64, 103, 110, 112]. Naturally, the accuracy of these algorithms depends on the task and datasets used. However, several reach the high eighties and nineties, showing their effectiveness on large-scale real-world datasets of fake reviews, fake news, and hoaxes. Detection algorithms are discussed in depth in Section 6.

Overall, this survey gives a comprehensive overview of the state of false information on the web and social media. Table 1 categorizes the research papers according to the platforms they study.

False Information

Intent   Knowledge

Misinformation   **Disinformation**   **Opinion-based**   **Fact-based**
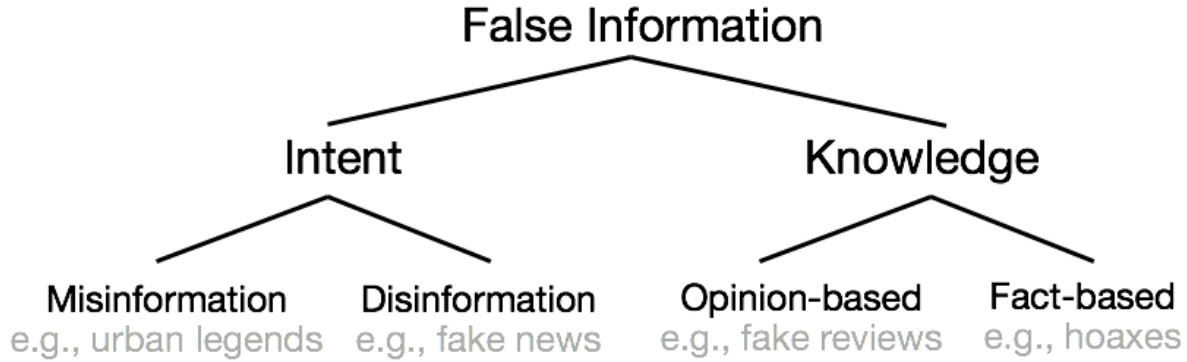e.g., urban legends   e.g., fake news   e.g., fake reviews   e.g., hoaxes

Fig. 1. Categorization of false information based on intent (i.e., is it spread with the intention to deceive or not) and knowledge (i.e., if there is a single ground

The remainder of the survey is organized as follows: Section 2 explains the two broad categories of false information, Section 3.2 discusses the mechanisms and rationale for the success of false information. Then, in Section 4, we describe the impact of false information. Section 5 elaborates on various characteristics of false information, and is followed by Section 6 which describes several algorithms for its detection.

2 TYPES OF FALSE INFORMATION

False information can be categorized based on its intent and knowledge content, as depicted in Figure 1. We discuss and detail this categorization here.

2.1 Categorization based on intent

False information can be classified based on the intent of the author, as *misinformation* and *disinformation* [25, 35]. By definition, misinformation is spread *without* the intent to deceive. Thus, common causes of misinformation include misrepresentation or distortion of an original piece of true information by an actor, due to lack of understanding, attention or even cognitive biases [24, 93]. These actors can then spread misinformation unwittingly to others via blogs, articles, comments, tweets, and so on. Note that readers can often simply have different interpretations and perception of the same piece of true information, leading to differences in how they communicate their understandings and in turn inform others' perception of the facts [4, 45].

Conversely, disinformation is spread *with* the intent to deceive [76]. Thus, understanding the motives for disinformation are much akin to understanding the motives for deception [25, 93]. Deception on the web occurs for many purposes, and for similar (though less interpersonal) reasons as in human interactions. The large potential audience leads most web disinformation campaigns to focus on swaying public opinion in one way or another, or driving online traffic to target websites to earn money by advertisements. One recent example is political disinformation spread during 2016 USA presidential elections [29, 38, 87], which even led to public shootings [27]. In this survey, we focus primarily on the technological aspects of disinformation, as the majority of research focuses around it.

2.2 Categorization based on knowledge

Under this categorization, false information is classified as either *opinion-based* or *fact-based* [101]. Opinion-based false information expresses individual opinion (whether honestly expressed or not) and describes cases in which there is no absolute ground truth. The creator of the opinion piece knowingly or unknowingly creates false
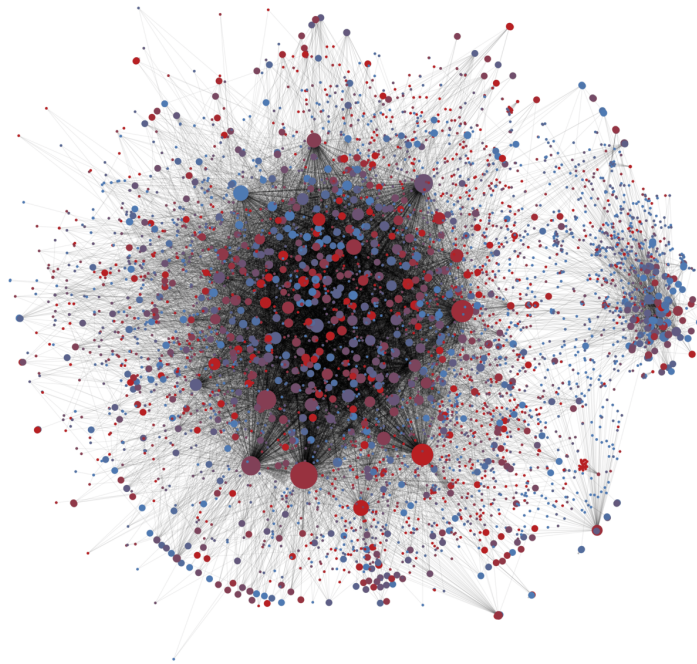
Fig. 2. False information spreads in social media via bots: Figure shows spread of #SB277 hashtag concerning a vaccination law. Red dots are likely bots and blue are likely humans. Figure reprinted with permission from [2].

opinions, potentially to influence the readers' opinion or decision. An example of false information that lies in this category is fake reviews of products on e-commerce websites, where people express their opinions about product quality. On the other hand, fact-based false information involves information which contradicts, fabricates, or conflates a single-valued ground truth information. The motive of this type of information is to make it harder for the reader to distinguish true from false information, and make them believe in the false version of the information [76]. This type of false information includes fake news, rumors, and fabricated hoaxes. There is significant research in both opinion-based and fact-based false information, and we will discuss both in this survey.

3 ACTORS AND RATIONALE OF SUCCESSFUL DECEPTION BY FALSE INFORMATION

This section describes the types of mechanisms used for spreading false information and the rationale behind their success.

3.1 Bad actors: bots and sockpuppets

Humans are susceptible to false information and spread false information [105]. However, the creation and spread of false information is complex, and fueled by the use of nefarious actors which act independently or on a large-scale using a network of social media bots. Both deceive readers by creating an illusion of consensus towards the false piece of information, for instance by echoing it multiple times or expressing direct support for it. These accounts aim to artificially *engineer* the virality of their content (e.g., by 'upvoting'/promoting

content in its early phase [109]) in order to spread posts with false information even faster and deeper than true information [11, 19, 30].

Lone-wolves operate by creating a handful of fake "sockpuppet" or "sybil" accounts and using them in coordination to reflect the same point of view, by writing similar reviews on e-commerce platforms or making similar comments on public forums. Lone-wolf operations using multiple accounts can be especially convincing as readers are typically not aware that a whole discussion is fabricated and actually originates from a single source. For instance, in online conversations, Kumar et al. [47] characterize this behavior by studying 61 million comments made by 2.1 million users across several discussion platforms. They found that while sockpuppets can be used with benign intention, sockpuppets with deceptive intentions are twice as common. Deceptive sockpuppets reply to each other with agreement and support, and are negative towards accounts that disagree. Moreover, these accounts hold central locations in the communication network, and are therefore in key spots to spread false content. Similarly, sybil accounts in communication and social networks are created to integrate themselves well into the network and prevent detection in order to increase their influence over others [113].

On a larger scale, social media botnets are used to spread false information. Bots, which are fake or compromised accounts controlled by a single individual or a program, are used to serve two main purposes: to send the same information to a large audience quickly, and to inflate the "social status" of certain users, both of which make false information to appear credible and legitimate [26, 85, 97]. Figure 10 visualizes an online Twitter conversation on a controversial topic (hashtag #SB277) showing overwhelming presence of bots (red nodes) engaging with humans (blue nodes) [2]. Bessi et al. [13] and Shao et al. [87] studied the use of bots in political campaigns and found that bot accounts are responsible for almost one-fifth of all Twitter political chatter, and that false information is more likely to be spread by bots than real users. Similarly, Nied et al. [66] found that 25% of false information tweets were generated by bots. A common strategy employed by bots is to target information towards more influential real users, who may sometimes get influenced and reshare the false message forward to a broader audience [13, 87]. In efforts to increase "social status", botnet operators offer services that provide fake followers by using their bots to follow paying customer accounts. Shah et al. [85] studied these services and found that they operate on "freemium" and "premium" models, where the former is made of compromised or real user accounts and the latter is comprised of fake or bot accounts. These two models operate quite distinctly—freemium fraud accounts create high-density cliques of opted-in accounts who trade follows amongst themselves, while premium fraud accounts create dense bipartite cores, i.e., one set of accounts follows the paying customers. This increases the apparent trustworthiness of the users, who can then be used to spread false information further.

In a recent study, Vosoughi et al. [105] analyzed over 126,000 false information cascades on Twitter over a period of 11 years and showed that humans were responsible for spread of false information on Twitter, not bots. Using the BotOrNot Twitter bot-detection tool developed by Davis et al [21], they identified the bot and non-bot accounts that engaged in false information spread. They found that on Twitter, humans, not bots, were responsible for spread of false information, as the bots were responsible for accelerating the spread of both true and false information roughly equally. Even after removing bot activity, false information was observed to spread farther, deeper, faster, and broader than true information. Further, they found that the non-bot accounts on Twitter that were responsible for spreading false information were newer, had fewer followers and followees, and were less active. While this is the case on Twitter, other platforms may behave differently, and the proliferation of nefarious actors in the creation and spread of false information is common.

Thus, using sockpuppets and botnets are used to engineer the spread of false information to massive numbers of real users on social media. These accounts operate using fake and computerized accounts to increase the visibility of false information and social status of accounts that spread it.
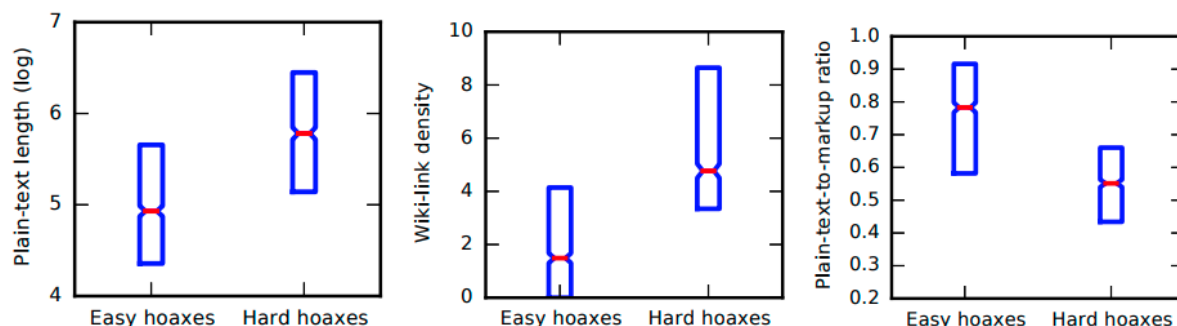
Fig. 3. Humans are unable to identify false information that is crafted to look genuine, as demonstrated in this case of hoaxes on Wikipedia. Reprinted with permission from [50].

3.2 Rationale of successful deception by false information

In the previous section, we discussed the multifaceted motives and spreading mechanisms used by those who publish false information. But what about the susceptibility of its consumers: do the readers tend to believe it, and if so, why?

*3.2.1 **Human inability to discern false information**.* False information would not have any influence if readers were able to tell that it is false. However, several research studies have conducted experiments to measure the ability of humans to detect false information including hoaxes, fake reviews, and fake news, and have shown that humans are not particularly good at discerning false from true information [50, 70, 74, 114]. We describe these studies in detail below.

To understand reader susceptibility, Kumar et al. [50] conducted an experiment with hoax articles created by hoaxsters on Wikipedia. They hired Amazon Mechanical Turk workers and showed them one hoax and one non-hoax article side-by-side, with the task to identify which one of the two articles was a hoax article without searching for information elsewhere. A total of 320 pairs of hoax and non-hoax articles were created and each pair was shown to 5 different workers. Humans correctly identified the hoax a mere 66% of times, only marginally higher than the random guessing baseline of 50%. They further studied the reasons for mistakes that workers made, shown in Figure 3, which compares several statistical properties of easily-identifiable and hard-to-identify hoaxes. They found that workers frequently misjudged long and well referenced hoax articles to be true, and short but true articles that lacked references to be hoaxes. In fact, even trained and trusted Wikipedia volunteers, called "patrollers," make the similar mistakes by approving long and well-referenced hoax articles for publication on Wikipedia instead of rejecting and deleting them. So, *if false information is purposefully created to look genuine, both trained and casual readers are deceived.* This indicates that humans give a lot of emphasis on the appearance of false information when judging its veracity.

In the domain of fake reviews, several of research studies have come to similar conclusions. Ott et al. [70] demonstrated that humans are not very good at discerning deceptive opinion spam from real reviews. As a compelling example, below are two TripAdvisor reviews (one real and one fake). Can you identify which one is fake?[3]

> *"I have stayed at many hotels traveling for both business and pleasure and I can honestly stay that The James is tops. The service at the hotel is first class. The rooms are modern and very comfortable. The*
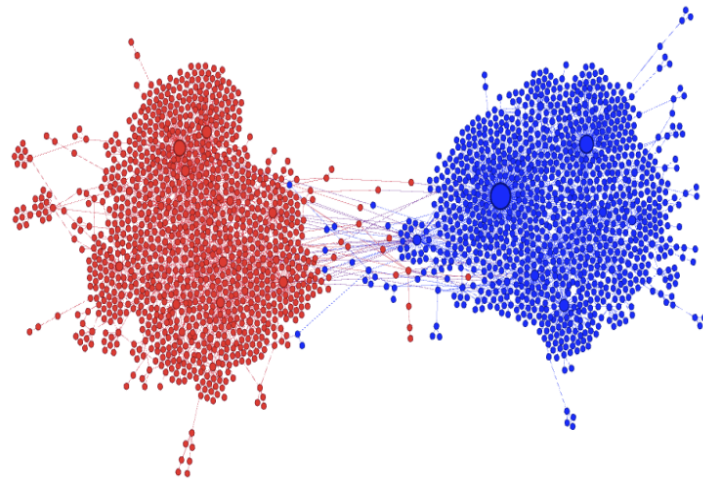
---

[3]The second review is fake.

Fig. 4. Social media platforms can produce echo-chambers, which lead to polarization and can encourage the spread of false information. Figure shown echo-chambers formation in retweet graph on controversial #beefban topoc. Reprinted with permission from [31].

> *location is perfect within walking distance to all of the great sights and restaurants. Highly recommend to both business travelers and couples."*

> *"My husband and I stayed at the James Chicago Hotel for our anniversary. This place is fantastic! We knew as soon as we arrived we made the right choice! The rooms are BEAUTIFUL and the staff very attentive and wonderful!! The area of the hotel is great, since I love to shop I couldn't ask for more!! We will definatly [sic] be back to Chicago and we will for sure be back to the James Chicago."*

The fake reviews were generated by Amazon Mechanical Turkers. Three humans were given a total of 160 reviews which contained both real and fake reviews, and workers had an accuracy between 53.1% and 61.9% in identifying the fake reviews, again showing that humans are poor judges of deception, and perform close to random.

For fake news, a similar recent study was conducted by Perez et al. [74]. They created a dataset of crowdsourced and crawled celebrity-oriented real and fake news, and gave 680 pieces of news (50% fake) to two humans to identify fake ones from them. They achieved an average accuracy of 70.5% in detecting made-up crowdsourced news, and 78.5% in detecting celebrity news.

More recently, with the advancement in deep learning, false information can be generated automatically. When fine tuned, this false information can be as deceptive as those created by humans. Yao et al. [114] created a deep neural network model that generates fake reviews for restaurants, by training on Yelp review data. Mechanical Turk workers were shown a set of 20 reviews for each restaurant, which contained between 0 and 5 machine generated reviews. The task of the workers was to identify which of the reviews were fake. A total of 600 sets of reviews were labeled, and the workers achieved a very low precision of 40.6% precision and 16.2% recall. Humans were able to identify fake reviews if they contained repetitive errors, but not when they had minor spelling or grammar mistakes.

Altogether, these four studies show that humans can easily be deceived into believing that false information is true when it is created intelligently to appear like true information, both manually or by machines.

3.2.2 **Formation of echo-chambers**. Given the advent of improved recommendation algorithms which promote personalized content for easy user access and exposure, social media platforms are often party to an "echo-chamber" effect [22]. This effect primarily refers to the self-selective polarizing effect of content where people immerse themselves in social circles in such a way that they are primarily exposed to content that agree with their beliefs. For example, a political liberal might friend more liberals on Facebook, thumbs-up liberal-minded content, and thus constantly be exposed to posts and news which aligns with his worldview. Figure 4 visualizes this echo-chamber effect on Twitter on a controversial topic of #beefban, where red and blue nodes represent users with opposing beliefs and edges represent who-retweets-whom, as shown by Garimella et al. [31]. Notice that both groups are mostly disconnected with few messages between nodes of different types. The echo-chamber effect in social networks is substantiated by Nikolov et al. [67] by demonstrating that the diversity of sources (links) clicked by users is significantly lower on social media platforms than in general search engines. Several studies have studied the effects and causes of echo-chambers. Quattrociocchi et al. [78] demonstrated that such resulting echo-chambers can serve to polarize the user's viewpoints by means of confirmation bias and lead to less diverse exposure and discussion between unaligned users. The resulting echo-chambers can contribute to the spread of false information by lowering the bar for critical fact-checking. Moreover, Trilling et al. [102] and Zajonc [116] posited that the perceived accuracy of false information increases linearly with the frequency of exposure of a participant to the same false information. This suggests that familiarity with repeatedly shared content (highly common and expected in echo-chambers) increases the perceived accuracy of the content, irrespective of its credibility. This calls for research on how to create effective techniques to break echo-chambers and slow down false information spread.

3.2.3 **Other reasons of successful deception**. Publishers of false information succeed at deceiving and spreading it by playing upon naiveté and biases of consumers. Flynn et al. [28] showed that prior belief in false information is rooted in the biased reasoning of the presented information. Two major factors that make consumers vulnerable or susceptible to believing false information are *confirmation bias* and *naive realism* [90]. Naive realism suggests consumers believe that they have the "true" perception of reality whereas disagreements or nonalignment of views is construed as the others' lack of rationality or cognizance [108]. Moreover, Nickerson et al. [65] characterized *confirmation bias*, or the tendency of consumers to seek or interpret evidence which confirms their pre-existing notions or beliefs. These biases lead consumers to look for and find meaning in pieces of information (no matter the veracity) which substantiate their own claims. For example, political liberals are prone to having more affinity towards posts promoting liberal viewpoints and condemning conservative ones, and vice versa. Furthermore, *social normative theory* suggests that sharing content aligned with the beliefs of their peers is attractive [10], in order to gain the acceptance or favor of their peers, regardless of its veracity. Alarmingly, Nyhan et al. [68] showed that even the presentation of corrections to false information by means of facts can actually further polarize ideological groups and *increase* their misperceptions.

Researchers have explored the psychological and demographics of information consumers and their propensity to believe in it. Pennycook et al. [73] investigated the psychological profiles to show a positive correlation between propensity for analytic thinking and the ability to discern false from real information, suggesting that false information often spreads due to poor analytical skills on the part of the consumer spreader. Additionally, inability to discern the publisher's original (possibly genuine) intentions can lead to consumer's misunderstanding of original information and lead to creation of misinformation [4]. Recent analysis of demographics by Allcott and Gentzkow [9] concluded that people who spend more time consuming media, people with higher education, and older people have a more accurate perception of information.

Overall, false information spread is orchestrated on large-scale by using fake social media accounts, such as bots and sockpuppets. Once their false message spreads and deceives readers, the readers themselves echo and spread the message. Several reasons lead to deception by false information. First, humans are unable to
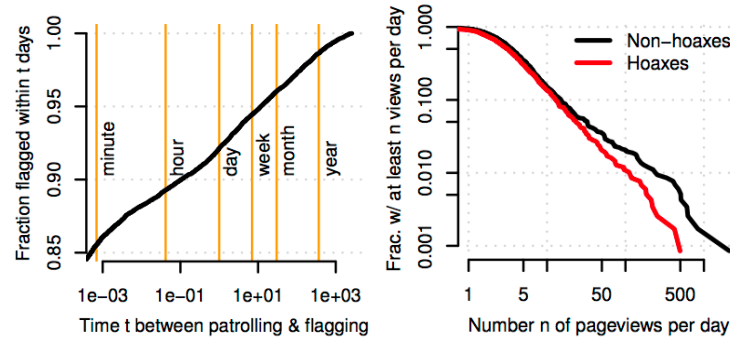
Fig. 5. False information is impactful as it (a) survives for a long time, and (b) is viewed by thousands of people. Reprinted with permission from [50].

distinguish false information from true ones when they come across them, and this is difficult even when they are trained to do so. Second, echo chambers are formed in social platforms, such that true and false information is spread among different groups of users, which do not interact with one another. And finally, human biases lead to increase in susceptibility, with some demographics (less educated and low consumers of media) being more likely to fall for false information.

4 IMPACT OF FALSE INFORMATION

Given that there are several factors that lead to deception by false information (Section 3.1), what is the impact of false information on its readers on web and social media? In the real world, false information has been shown to have significant impact on the stock market [17], hampering response during natural disasters [32], and terroristic activity [27, 96]. On web and social media, the impact is measured as the engagement it produces via its readers, using statistics such as number of reads, number of days it survived without being removed, or number of people it reached via reshares. Several research studies have been conducted to measure the impact of hoaxes, fake reviews, and fake news. Frigerri et al. [30] studied the spread of rumors on Facebook, Kumar et al. [50] measured the impact of hoax articles on Wikipedia, and Silverman [92] analyzed the engagement of fake election news articles on Facebook. We discuss these studies to measure the impact of false information on web platforms.

False information spreads far and wide on social media because there is an average delay of 12 hours between start of false information spread and that of its debunking information [86, 121]. False information spreads rapidly during its starting phase—an unverified and not yet debunked rumor has high potential of becoming viral [121]. As a result, rumors with the possibility of being true start to spread, sometimes even by reputed news organizations [91].

On Wikipedia, Kumar et al. [50] measured impact of hoaxes in terms of their viewcount, number of days they survived before they are deleted, and their spread across the web. Figure 5 shows the distributions for the first two statistics. Figure 5(a) shows the distribution of the time it takes from when the article is created and approved ('patrolled') till the time it is identified as a hoax ('flagged'). It shows that while 90% of hoax articles are identified immediately within an hour of being approved, about 1% of hoaxes that are well-written hoaxes survive for over one year without being detected. However, survival is not enough for a hoax article to be successful; it must be viewed as well. Figure 5(b) plots the counter-cumulative distribution of average view count of hoaxes that survive for at least a week and their equivalent non-hoaxes. On average, hoaxes are viewed less frequently than non-hoaxes (median 3 views per day vs 3.5 views per day), but a non-negligible 1% of hoaxes are viewed at least 100 times a day. Finally, the impact of hoaxes is measured in terms of spread over the web, by counting the links

that were clicked by readers to reach the hoax article. For this, 5 months of Wikipedia server logs were used. The results were alarming—at least 5 distinct links were clicked from across the web for 7% of hoaxes, and on average, each hoax had 1.1 such links. This traffic was observed from search engines, social networks such as Facebook and Twitter, and from within Wikipedia itself. Overall, this analysis shows that while most hoax articles are ineffective, a small fraction of hoax articles on Wikipedia is highly impactful.

Buzzfeed news analyzed highly impactful fake political news on the web. They analyzed both true and false election-related stories with the highest engagement on Facebook during the 2016 US Presidential election [92]. Engagement was measured as the total number of shares, reactions, and comments on the Facebook story. They analyzed the 20 top-performing false election stories generated by fake websites and blogs, and compared them to the 20 top-performing true election stories from major news websites, like New York Times, Washington Post, and others. The fake news stories got a total of 8,711,000 engagements, significantly higher than the 7,367,000 engagements of the real news stories. As this analysis was restricted to top stories, a complete analysis of all news stories may reveal a different picture. Prior to this study, Gupta et al. [32] studied the spread of eight fake images on Twitter during Hurricane Sandy, and found that that fake images were shared almost twice as much as real images.

On a larger scale, Frigerri et al. [30] conducted a comprehensive study of the spread of false and real information on Facebook. They collected 4,761 rumors from *snopes.com*, which is a website that catalogues popular stories on social media and checks their veracity. In their dataset, 45% of stories were fake, 26% were "true" (i.e., not a fake story), and the rest had intermediate truth values. They analyzed the rumor cascades propagating as photos on Facebook during July and August 2013. Each cascade was identified as a tree of reshares starting from the original post of the photo, whenever a link to a valid snopes article was posted as a comment to the original photo or one of its reshares. A total of 16,672 such cascades were identified, with 62,497,651 shares, showing the large visibility false rumors can have. Surprisingly, they found that false information cascades were deeper, as there were more reshares at greater depths than the reference cascades. At lower depth, i.e., closer to the original photo post, the reference cascades have more reshares—about 20% reference cascades have depth of at least two, compared to 10% of false information cascades. But the reference cascades die very soon, while false information cascades run deeper. About 3% of false cascades have depth of at least 10 reshares, while less than 1% of reference cascades have the same property. The difference increases in magnitude as the depth of the cascade increases. This study shows the large reach of false information on social media, fueled by its highly contagious nature.

Recently, the largest study of spread of over 126,000 rumors on Twitter over a period of 11 years was conducted by Vosoughi et al. [105]. The authors took the set of false information cascade identified by various independent fact-checking agencies and traced their spread from their very beginning. This was done by identifying cascades that contained a link to any of the agencies. For comparison, they also considered cascades of verified true information linking to these agencies. Compared to true information, tweets containing false information spread significantly farther (more number of users retweeted), faster (more number of retweets in a shorter time), deeper (more number of retweet hops), and more broadly (more number of users at some retweet depth). This was observed in all categories of false information, such as politics, urban legend, science, business, and others, with politics as the biggest category of false information. In fact, they found that the top 1% of false tweets reached over 1,000 users, which true information tweets rarely did. False information reached more number of people than truth at every cascade depth, which was aided by its virality, showing that it was spread by multiple people in a peer-to-peer manner, instead of a few accounts simply broadcasting it. Moreover, false information was six times faster in reaching the same number of people as true information did. Thus, this study showed the widespread reach and impact of false information in Twittersphere.

Overall, impact of false information on the web is measured using engagement statistics such as view count, share count, and more. Research has shown that while most false information is not effective, a small fraction

Feature category	Opinion-based false information (fake reviews)	Fact-based false information (false news and hoaxes)
Text	Harris et al. [33], Jindal et al. [43], Li et al. [51] , Lin et al. [55], Mukherjee et al. [61, 62] , Ott et al. [69, 70], Rayana et al. [79], Yao et al. [114]	Gupta et al. [32], Horne et al. [37], Howard et al. [38], Kumar et al. [50], Mitra et al. [59, 60], Perez et al. [74], Qazvinian et al. [77], Rubin et al. [80], Silverman et al. [91], Wang et al. [107], Yang et al. [112], Zeng et al. [118]
User	Hooi et al. [36], Kumar et al. [48], Li et al. [51], Minnich et al. [58], Mukherjee et al. [61], Rayana et al. [79], Shah et al. [84]	Bessi et al. [13], Davis et al. [21], Gupta et al. [32] Jin et al. [41], Kumar et al. [50], Mendoza et al. [57] Nied et al. [66], Shao et al. [86, 87] , Tacchini et al. [100], Vosoughi et al. [105], Yang et al. [112]
Graph	Lin et al. [55], Minnich et al. [58], Pandit et al. [71], Rayana et al. [79], Shah et al. [83]	Bessi et al. [13], Davis et al. [21], Friggeri et al. [30], Kumar et al. [50], Mendoza et al. [57], Nied et al. [66], Qazvinian et al. [77], Starbird et al. [95] Subrahmanian et al. [97], Vosoughi et al. [105]
Rating score	Beutel et al. [15], Harris et al. [33], Hooi et al. [36], Kumar et al. [48], Li et al. [51], Luca et al. [56], Minnich et al. [58], Mukherjee et al. [61, 62], Rayana et al. [79], Shah et al. [84], Ye et al. [115]	Not applicable
Time	Hooi et al. [36], Li et al. [52, 53], Minnich et al. [58], Mukherjee et al. [61], Rayana et al. [79], Shah et al. [84], Xie et al. [111], Ye et al. [115]	Davis et al. [21], Del et al. [22], Friggeri et al. [30] Shao et al. [87], Vosoughi et al. [105], Zannettou et al. [117], Zeng et al. [118]
Propagation	Not applicable	Friggeri et al. [30], Jin et al. [41], Shao et al. [86, 87], Silverman et al. [91, 92], Vosoughi et al. [105], Yang et al. [112], Zannettou et al. [117], Zeng et al. [118], Zubiaga et al. [121]

Table 2. This table categorizes research based on the characteristics and features of false information analyzed.

(typically 1%) is highly impactful, most popular false information pieces attract more attention than real information, and false information spreads widely and quickly across the web and reaches a large population on social media.

5 CHARACTERISTICS OF FALSE INFORMATION

In this section, we describe several characteristics of false information that act as "tell-tale" signs for discernment from true information. These characteristics are based on textual content, time, ratings, graph structure, creator
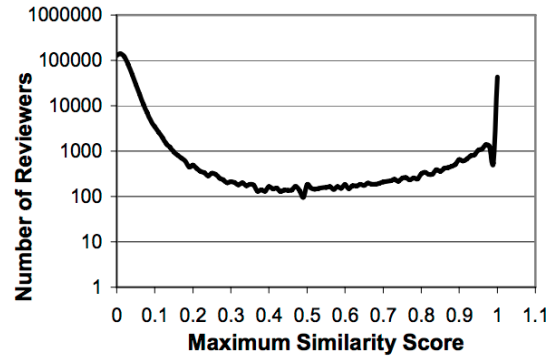
Fig. 6. Plot showing maximum similarity of reviews between two reviewers, showing that 6% reviewers have at least one identical review to someone else's review. Reprinted with permission from [43].

properties and more. We separately discuss characteristics of opinion-based and fact-based false information, along these axes. Table 2 categorizes the research papers according to the features that they study, as it is one of the most common approaches to approach the problem of false information. The types of features can broadly be grouped as text, user, graph, rating score, time, and propagation-based features. We will discuss these in detail in this section.

5.1 Opinion-based false information

Below, we discuss several discovered properties of fake ratings and e-commerce reviews. We categorize properties pertaining to i) text, ii) ratings, iii) time, iv) graph structure and v) other attributes separately, as these aspects have both individually and in conjunction received considerable attention from researchers.

5.1.1 Textual characteristics. Since most reviews include textual content, researchers have extensively studied textual and linguistic features for discerning review fraud. Several works have posited that review fraudsters minimize effort by repeating the same reviews. Jindal et al. [43] provided the first well-known characterizations of review fraud, in which the authors characterized duplicate reviews (according to Jaccard similarity) across Amazon data as cases of fraud. The authors showed that many of these fraudulent duplicate reviews were from the same user on different products, rather than different users on the same product or different products. Figure 6 shows the distribution of maximum similarity between two reviewers' reviews. At the higher similarity end, 6% of the reviewers with more than one review have a maximum similarity score of 1, which is a sudden jump indicating that many reviewers copy reviews. Furthermore, Sandulescu et al. [82] showed that many review fraudsters adjust their reviews slightly so as not to post near or exactly similar reviews and be easily caught—instead, these sophisticated fraudsters tend to post semantically similar text (i.e. instead of duplicating "the hotel room had an excellent view," the fraudster might post "the hotel room had a superb view" instead).

Researchers have also focused more on the linguistic features of deceptive reviews, such as using stylistic analysis (number of words, characters, etc.), lexical analysis (number of verbs, nouns, etc.), psycholinguistic analysis (LIWC [72]), and sentiment analysis (emotion, sentiment, etc.). Mukherjee et al. [62] showed that fake reviews were shorter than real reviews, and Ott et al. [70] found that imaginative "faked" writing is typically more exaggerated and consists of more verbs, adverbs, pronouns and pre-determiners. Furthermore, deceptive text tends to have an increased focus on aspects external to the venue being reviewed (more emphasis on family, vacation, business, etc.) [70]. Looking at negative reviews, Ott el at. [69] found that fake negative review writers exaggerate too negatively, including words which communicated negative emotion far more than normal reviews
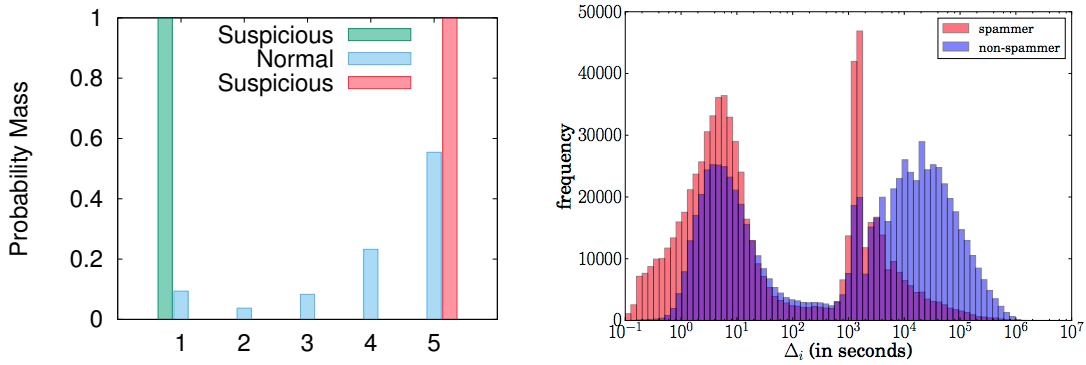
Fig. 7. (a) Aggregate e-commerce rating behavior typically follows the *J*-shaped curve in blue, whereas review spammers commonly have strongly positively or negatively-biased rating distributions like those in green and red [84]. (b) Fraudulent and non-fraudulent users have bimodal rating distribution [53]. Figures reprinted with permission from [84] and [53].

(terrible, disappointed, etc.). Furthermore, fake reviews eschew the use of pronouns such as "I," perhaps in order to distance themselves from the negative sentiments. Similar observations were made by Li et al. [54] on fake reviews generated by domain experts. Finally, Harris [33] demonstrated that deceptive opinion spam tends to be on average less readable than truthful reviews (measured by Average Readability Index), and is also more polarized and sentimental than those reviews, supporting previous results.

*5.1.2 **Ratings characteristics** .* Many e-commerce sites disallow users from giving feedback without giving an associated numerical rating. The rating is typically a 5-star system (1 representing the worst possible rating, and 5 representing the best), and is employed by numerous major online marketplaces including Amazon, eBay, Flipkart, and more. Prior work in review fraud has shown that those who engage in spreading fake e-commerce reviews also typically have skewed rating distributions [15, 36, 79, 84] which are not typical of real users who share non-uniform opinions over many products. Figure 7 shows an example from Shah et al. [84], comparing aggregate (dataset-wide) rating habits from the Flipkart platform with two common, naive fraudster rating habits depicting very positive and negative raters. Some fraudulent reviewers give only positive ratings as they are created in order to inflate ratings for customer products, whereas other such reviewers give only negative ratings as they intend to slander competitors' products. Further, Kumar et al. [48] recently showed that fraudulent review writers are typically unfair, in that they give "unreliable" rating scores that differ largely from the product's average score. Furthermore, these fraudulent writers often give high ratings to products that otherwise receive highly negative ratings from fair users.

*5.1.3 **Temporal characteristics**.* Fraudulent review writers typically give reviews in "lockstep," or at the same/similar times. The rationale is similar to that for dense subgraph connectivity—the review writer's accounts are often controlled by scripts, and are thus temporally synchronized in short windows. A number of papers have leveraged the distribution of interarrival times (IATs) between each user's successive ratings/reviews to detect review spammers. Shah et al. [84], Hooi et al. [36], and Ye et al. [115] showed that in e-commerce websites, spammers are often characteristic of very short IATs (on the order of seconds or minutes) between subsequent ratings, unlike typical users who would rate sporadically and likely only upon making and receiving a purchase. Xie et al. [111] substantiated these findings, with particular emphasis on singleton review spammer attacks. Further, Li et al. [52] and Minnich et al. [58] showed that many fraudulent check-ins/reviews in such networks occur with short, but moreover "infeasible" IATs between check-ins. Since users must be physically present at a
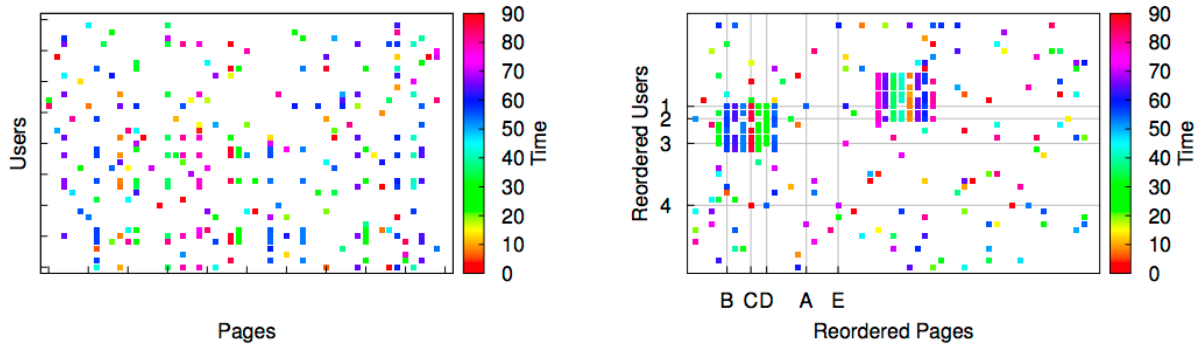
Fig. 8. Fraudulent reviewers often operate in coordinated or "lock-step" manner, which can be represented as temporally coherent dense blocks in the underlying graph adjacency matrix. Reprinted with permission from [14].

location to be allowed to check-in and leave a review for a location, reviews at far-away places in very short IATs are a notable distinguishing characteristic of fraud.

In addition, more recent research by Li et al. [53] surprisingly found that the posting rates of both fraudulent and non-fraudulent users is bimodal—some reviews are written in a short time bursts, while some others with more time between consecutive reviews. Fraudulent users are still more bursty than non-fraudulent users, as the latter have the tendency to be more active after a period of inaction to summarize their recent experiences.

*5.1.4 **Graph-based characteristics**.* Several works show that dense subgraphs produced by coordinated or "lock-step" behavior in the underlying connections of the social (in this case, review) graph are associated with fraudulent behavior [16, 71, 83]. Figure 8 demonstrates this pattern in page-likes on Facebook [16]. Alternatively, other works look at the local network structure of the users instead of global structure. For example, Lin et al. [55] showed that for review platforms where multiple ratings/reviews can be given to the same product, review fraudsters often repeatedly post to the same product instead of diversifying like a real reviewer.

Studying group structure, Mukherjee et al. [61] showed that ratios for review group (defined as a set of reviewers who have reviewed at least k common products) size to the total number of reviewers for an associated product tend to be significantly higher in fraudster groups than real users. This is because many products (especially bad ones) have ratings/reviews almost entirely given by fraudsters, whereas this case is uncommon for real reviewers. Furthermore, fraudster groups tend to have larger group size and higher support count (in that they share a large number of target products)—these features essentially reflect the size and density of the group's subgraph.

Incorporating time component, Beutel et al. [16] extended the group definition beyond graph connections by incorporating temporal closeness, and shows that group fraud (e.g., bots coordinating to post fake reviews) is temporally coherent as well and forms bipartite cores in a rearranged user-page bipartite network (Figure 8). The existence of large temporally-coherent bipartite cores is highly suggestive of fraud.

Overall, opinion-based false information tends to be shorter, more exaggerated, and has more extreme ratings (1-stars and 5-stars). The fraudsters that create this false information give several ratings in a short time period ('bursty') and operate in a coordinated fashion ('lockstep').

5.2 Characteristics of fact-based false information

In this section, we discuss false information concerning facts with a single-valued truth. This is different from information that may vary by someone's own opinion, e.g., their opinion about a particular product on Amazon. Specifically, we discuss here the characteristics of hoaxes, rumors, and fake news.

5.2.1 Textual characteristics. There is a vast literature that studies fake news in social media. False information in the form of fake news is created in such a way to invoke interest and/or be believable to consumers. Various strategies may be used to deceive these consumers. Silverman [91] found that about 13% of over 1600 news articles had incoherent headline and content body, for example, by using declarative headlines paired with bodies which are skeptical about the veracity of the information.

In a recent paper, Horne and Adali [37] studied the textual characteristics of fake news using several sources of data: Buzzfeed fake news analysis articles [92], and articles from well known satire and fake news agencies (e.g., The Onion, Ending the Fed, and others). Reputed journalistic websites were used for comparison. The authors find interesting relations by comparing fake, satirical, and real news. Below are two news article titles, one of which is fake. Can you identify the fake one?[4]

> 1. *BREAKING BOMBSHELL: NYPD Blows Whistle on New Hillary Emails: Money Laundering, Sex Crimes with Children, Child Exploitation, Pay to Play, Perjury*
>
> 2. *Preexisting Conditions and Republican Plans to Replace Obamacare*

Fake news tends to pack the main claim of the article into its title. The titles are longer but use fewer stopwords and more proper nouns and verb phrases, meaning that the creators tend to put as much information in the title as possible. The words used in the title are smaller and capitalized more often, to generate emphasis. Not surprisingly, titles of fake news and satire are very similar. In terms of the body content, fake news articles are short, repetitive, and less informative (fewer nouns and analytical words). They contain fewer technical words, more smaller words, and are generally easier to read. This allows the reader to skip reading the entire article, and instead just take information away from the title itself, which may be disparate from the rest of the content of the article. Interestingly, Rubin et al. [80] studied satire news separately from the viewpoint of misleading readers into believing it is true, and also found that satirical articles pack a lot of information in single sentences. Thus, fake news articles are more similar to satirical ones than to real news—the bodies are less wordy and contain fewer nouns, technical and analytical words. In addition, Perez et al. [74] also analyzed the textual properties of fake news using two datasets—one generated by Amazon Mechanical Turk workers and other one scraped on celebrity rumors from gossip websites. They found that fake news contains more social and positive words, is more certain, focuses more on present and future actions by using more verbs and time words.

But do people discuss false information differently from true information? To answer this, Mitra et a. [60] recently analyzed the language of how people discuss true and false information pieces using tweets of 1400 events. These events were part of their CREDBANK dataset, which used crowdsourcing to label ground truth credibility judgments [59]. Using LIWC categories [72], they found that discussions around false information are marked with increased use of more confusion, disbelief, and hedging words which indicates skepticism among readers. Surprisingly, they found while more agreement words signaled high credibility, more positive sentiment words are associated with low credibility events. The latter is because it includes words like 'ha', 'grins', 'joking' are positive sentiments, but instead mean mockery. Their findings show that in addition to the text of the tweet itself, its surrounding discussion give important information to identify false information.

Hoaxes have similar textual properties as rumors. Kumar et al. [50] compared the content of hoax articles and non-hoax articles. They found that hoaxes were surprisingly longer compared to non-hoax articles (Figure 9(a)), but they contained far fewer web and internal Wikipedia references (Figure 9(b)). This indicated that hoaxsters
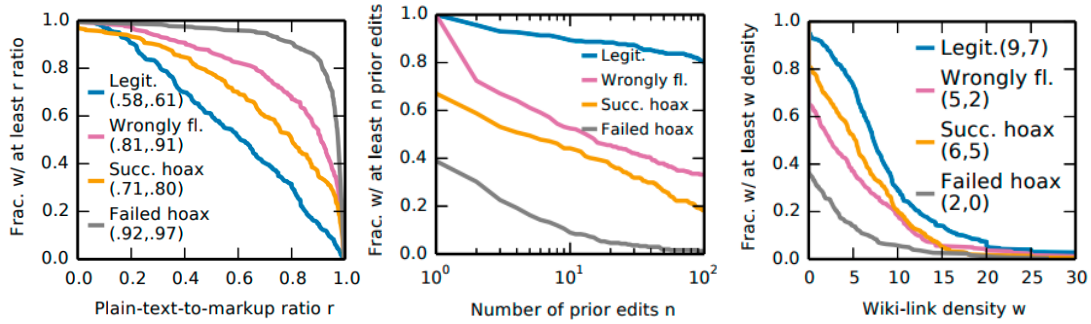
---

[4]First headline is fake.

Fig. 9. Hoax articles (a) have lots of text, (b) fewer references, and (c) are created by newer accounts. Figure reprinted with permission from [50].

tried to give more information to appear more genuine, though they did not have sufficient sources to substantiate their claims.

*5.2.2 **User characteristics***. Several studies have shown that the characteristics of *creators* of false information are different from those of true information creators. Kumar et al. [50] found that the creators of hoaxes have typically more recently registered accounts and less editing experience (Figure 9(c)), suggesting the use of "throw-away" accounts. Surprisingly, non-hoax articles that are wrongly assumed to be hoaxes were also created by similar editors, meaning that they lack the skills to create well-written articles, which leads to others believing that the article is a hoax.

In cases of rumors, Bessi et al. [13] studied over 20.7 million tweets related to US presidential election, and identified users involved in tweeting as bots or honest users using a classification tool produced by Davis et al. [21]. Their analysis found that about one-fifth of content created and spread was by bots, showing that rumors are spread by automated accounts in short-bursts of time. Shao et al. [87] came to similar conclusions in their experiments.

*5.2.3 **Network characteristics***. Rumors and hoaxes can be related to other information in terms of what they say about others and what others say about it. Kumar et al. [50] quantified this for hoaxes on Wikipedia by measuring the connectedness of the different Wikipedia articles referenced in the hoax article. Intuitively, high connectedness indicates interrelated and coherent references. The authors computed the clustering coefficient of the local hyperlink network of the article, i.e., the average clustering coefficient of the subnetwork induced by the articles referenced by the article. They found that hoax information has fewer references and significantly lower clustering coefficient compared to non-hoax articles. This suggests that references in hoaxes are added primarily to appear genuine, instead of adding them by need as legitimate writers do.

Network characteristics of rumors are studied by analyzing the network of users that spread them and by creating co-occurrence networks out of false information tweets—these contain nodes of one or more types, such as URLs, domains, user accounts or hashtags, and use edges to represent the number of times they are mentioned in the same tweet together. Using the user-user network, Subrahmanian et al. [97] found that some bot accounts that spread false information are close to each other and appear as groups in Twitter's follower-followee network, with significant overlap between their followers and followees. Moreover, Bessi et al. [13] conducted a k-core analysis of this follower-followee network and found that the fraction of bots increases steadily in higher cores, suggests that bots become increasingly central in the rebroadcasting network. Using the co-occurrence network,
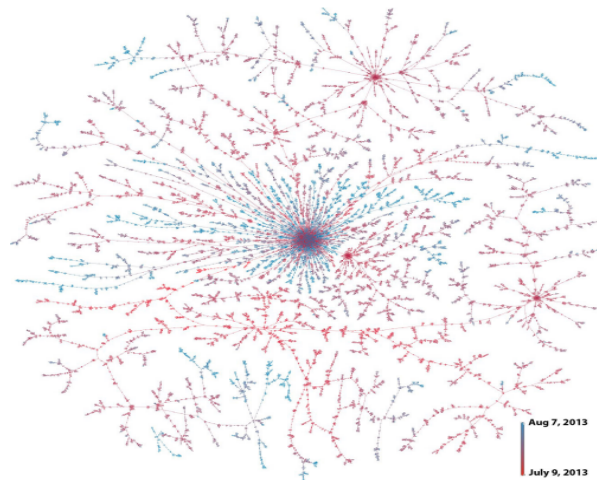
Fig. 10. Cascade of reshares of a Cabela's sporting goods store receipt attributing addition sales tax to "Obamacare". The coloring is from early (red) to late (blue). Reprinted with permission from [30].

Starbird [95] found that alternate media (false news) domains form tightly connected clusters, meaning that many users mention these domains together in their false information tweets.

5.2.4 ***Propagation characteristics***. The spread of false information in social media makes it highly impactful. Several research studies have shown that only a small fraction of users are responsible for most of the spread, instead of being akin to a grass-roots movement. Gupta et al. [32] found that the top 30 users contributed towards 90% of the retweets of fake images during hurricane Sandy on Twitter. Shao et al. [86] came to a similar conclusion in their study of about 1.3 million rumor tweets as well. Their analysis suggested that fake news spread was mostly dominated by a handful of very active users, whereas fact-checking of rumors was a more grass-roots activity with more conversation, and therefore slower. This suggests that repetition and perseverance play an important role in the spread of false information. Since people tend to spread unverified claims [91, 121], making false information believable may not be as important as persistently spreading it.

When false information spreads in social platforms, it spreads deeper compared to real news. In their study of rumor reshares on Facebook, Frigerri et al. [30] concluded that false information reshare cascades spread much deeper compared to that of true reference cascades. In other words, they are more likely to be reshared at greater depth and thus reach more people. One such reshare cascade is shown in Figure 10, with cascades colored by time. Additionally, Zeng et al. [118] showed that information related to rumors, both supportive and denying, spread faster than non-rumors. Simulations conducted by Doerr et al. [23] on realistic spread of simple rumors, on several graphs having the structure of existing large social networks, showed that even a rumor started by a random node on average reaches 45.6 million of the total of 51.2 million members within only eight rounds of communication. This is corroborated by the bursty behavior of rumors shown in several other research studies [92, 121].

Several researchers have shown that false information spreads quickly, especially during its early stage. Zubiaga et al. [121] studied the entire lifecycle of true and false information as it spreads through social media, both before and after its veracity is checked. They collected 330 rumor threads with 4,842 tweets of nine popular cases, such as Charlie Hebdo shooting, and Michael Essien contracting Ebola. Journalists then annotated the discussion threads of these rumors to quantify support expressed in tweets, i.e., their level of certainty (level of confidence

indicated by the tweet) and evidence (whether the tweet substantiates the rumor). They found that the spread of false information occurs largely before it is even debunked. Tweets that supported unverified claims generated the most retweets, sparked by sudden bursts of retweets even during the first few minutes, with interest in the rumor decreasing substantially after its veracity is checked. During the initial spread of information, all users including normal users as well as reputed ones affiliated with news organizations, tend to tweet with a bias towards supporting unverified claims as opposed to denying them, irrespective of whether the information is later confirmed or denied. Silverman [91] corroborates this finding. The level of certainty of tweets tends to remain the same before and after information is fact-checked, but users give more evidence before the rumor is fact-checked and less later on. These findings together further evidence the virality of popular false information during its initial phase.Further, Vosoughi et al. [105] also showed that tweets about false information spread significantly farther, deeper, faster, and broader than those about true information. This was observed for all categories of false information, such as politics, urban legend, science, business, and more.

While the above studies focus on spread of (false) information on a single platform, recent studies by Zannettou et al. [117] and Albright [7, 8] mapped the false information ecosystem across social media platforms. Zannettou et al. [117] studied the temporal relation between same information piece appearing on Twitter, Reddit, and 4chan platforms. They found that false information pieces are more likely to spread across platforms (18% appear on multiple platforms) compared to true information (11%). Moreover, false information appears faster across platforms than legitimate ones, and seems to 'flow' from one to another, with Reddit to Twitter to 4chan being the most common route. This spread across platforms is dangerous—Albright [8] studied the logs seven false information websites, and found that a whooping 60% of incoming traffic was from Facebook and Twitter, and rest were from emails, search engines, messaging, or direct visits. To study how these platforms connect to one another, Albright [7] crawled 117 false information websites and created a hyperlink network of domains that these websites refer to. He found that right-wing news websites link highly to other similar websites, thus supporting each other. Very surprisingly, YouTube was the most linked website overall, suggesting high use of multimedia content in conveying false information messages.

Thus, these studies have found that false information tends to propagate deeper and faster than true information, especially during the early stages of the false information. This happens on a single as well as across multiple platforms, and a handful of users are primarily responsible for this spread.

5.2.5   *Debunking characteristics.* Once false information spreads, attempts are made to debunk it and limit its spread. Recent research has shown that there is a significant time delay between the spread and its debunking. Zubiaga et al. [121] found that true information tends to be resolved faster than false information, which tends to take about 14 hours to be debunked. Shao et al. [86] came to a similar conclusion—they found a delay of 10–20 hours between the start of a rumor and sharing of its fact-checking contents.

But once debunking information reaches the rumor spreaders, do they stop spreading it or does it 'back-fire', as observed in in-lab settings [68] where corrections led to an *increase* in misperception? Several empirical studies on web-based false information suggest that debunking rumors is in fact effective, and people start deleting and questioning the rumor when presented with corrective information. Frigerri et al. [30] studied the spread of thousands of rumor reshare cascades on Facebook, and found that false information is more likely to be linked to debunking articles than true information. Moreover, once it is linked, it leads to a 4.4 times increase in deletion probability of false information than when it is not, and the probability is even higher if the link is made shortly after the post is created. Moreover, Zubiaga et al. [121] found that there are more tweets denying a rumor than supporting it after it is debunked, while prior to debunking, more tweets support the rumor. Furthermore, Vosoughi et al. [105] showed that there is a striking difference between replies on tweet containing false information than those containing true information—while people express fear, disgust, and surprise in

False information detection algorithms

Feature
engineering

e.g., using text,
time, user, metadata

Graph
algorithms

e.g., lockstep, dense
block detection

Modeling
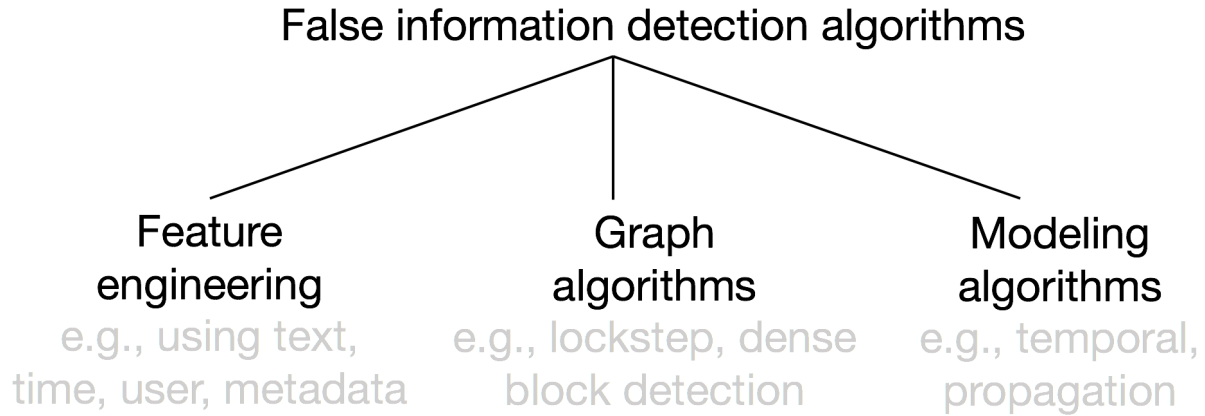algorithms

e.g., temporal,
propagation

Fig. 11. Algorithms to detect false information, both opinion-based and text-based, can be broadly classified into (a) feature engineering, (b) graph algorithms, and (c) modeling algorithms.

replies, true information generates anticipation, sadness, joy, and trust. These differences can potentially be used to create early detection and debunking tools.

Overall, research on characterization of fact-based false information has shown that it tends to be longer, generates more disbelief and confusion during discussions, is created by newer and less experienced accounts that are tightly connected to each other, spreads deeper and faster in one and across multiple platforms, and gets deleted when debunking information spreads.

6 DETECTION OF FALSE INFORMATION

In the previous section, we discussed a number of tell-tale signs and often-found characteristics of opinion-based and fact-based false information. In this section, we complement this information by discussing a number of approaches that researchers have employed to actually detect false information and those who spread it.

Algorithms to identify false information can be broadly categorized into three categories: feature engineering-based, graph-based, and modeling-based, as shown in Figure 11. The majority of algorithms are *feature-based*, in that they rely on developing efficient features that individually or jointly are able to distinguish between true and false information. These features are developed from the characterization analyzes that show the differences in properties of the two classes. These differences are then characterized by intelligently designed features. While we go into the details of some key research in feature-based detection, other papers that use features as described in Section 5 can directly be applied for detecting false information as well. Alternatively, *graph-based* algorithms rely on identifying false information by targeting groups of users (spreaders) with unlikely high, lock-step coordination boosting a certain story (e.g., a botnet retweeting the same article in near-identical time). These algorithms try to identify dense blocks of activity in an underlying adjacency matrix. While these algorithms may be able to identify large-scale coordinated activity, small-scale or lone-wolf attacks are unlikely to be caught since the algorithms primarily focus on the largest dense blocks. Finally, *modeling-based* algorithms work by creating information propagation models that emulate the empirical observation of edges and information spread. The intuition behind these algorithms is that since most information is true, it likely spreads a similar or unique way. Thus, emulating this mode of information spread can pinpoint false information spread as anomalies which can then be verified and removed. In this section, we will look broadly at approaches belonging to these three

Algorithm category	Opinion-based false information (fake reviews)	Fact-based false information (false news and hoaxes)
Feature-based	Harris et al. [33], Jindal et al. [43], Li et al. [51, 52, 54], Lin et al. [55], Minnich et al. [58], Mukherjee et al. [61, 62], Ott et al. [69, 70], Sandulescu et al. [82]	Gupta et al. [32], Horne et al. [37], Kumar et al. [50], Perez et al. [74], Qazvinian et al. [77], Rubin et al. [80], Wang et al. [107], Yang et al. [112]
Graph-based	Akoglu et al. [5], Hooi et al. [36], Kumar et al. [48], Rayana et al. [79], Shah et al. [84], Wang et al. [106]	Jin et al. [42], Ruchansky et al. [81], Shu et al. [89], Tacchini et al. [100]
Modeling-based	Temporal: Xie et al. [111], Ye et al. [115] Sentiment: Li et al. [54]	Propagation: Acemoglu et al. [3], Budak et al. [18], Del et al. [22], Doerr et al. [23], Jin et al. [41], Nguyen et al. [64], Tripathy et al. [103], Wu et al. [110], Zhao et al. [119]

Table 3. This table categorizes research research based on the type of false information detection algorithm.

classes for opinion-based and knowledge-based false information detection. Table 3 categorizes the papers that develop detection algorithms into these three categories.

The task of finding false information is one rife with challenges [98]. One of the major challenges arises from the imbalance in the population of two classes, false and true information In almost all cases, false information comprises only of a small fraction (less than 10%) of the total number of instances. Moreover, false information is masqueraded to seem like truth, making it harder to identify. And finally, obtaining labels for false information is a challenging task. Traditionally, these are obtained manually by experts, trained volunteers, or Amazon Mechanical Turk workers. The process requires considerable manual effort, and the evaluators are potentially unable to identify all misinformation that they come across. Any algorithm that is developed to identify false information must seek to address these challenges.

6.1 Detection of opinion-based false information

Here we look at the algorithms that have been developed in literature to identify opinion-based false information. Specifically, we look at the research on identifying fake reviews in online platforms using text, time, and graph algorithms. Text-based algorithms primarily convert the textual information into a huge feature vector and feed that vector into supervised learning models to identify duplicate and fake reviews. Graph-based algorithms leverage the user-review-product graph to propagate beliefs and to jointly model 'trustworthiness' of users, reviews, and products. Time-based algorithms employ time-series modeling and co-clustering along with feature engineering. We will elaborate on these algorithms in the next few subsections.

6.1.1   Feature-based detection. As text is the primary source to convey (false) information on web platforms, it is one of the most widely studied component for fake review detection. Algorithms in this domain are based on *feature engineering*, detecting *duplicate reviews*, or a *combination* of the two. We primarily focus on text-based detection, as other features, such as user, graph, score, and time, are usually used in conjunction with text features in this task.

Several algorithms have been developed to efficiently identify duplicate reviews, with the notion that fraudsters give identical or near-identical reviews while genuine reviewers give more unique reviews. Jindal et al. [43]

studied three major types of duplicates: different users reviewing the same product, same user reviewing different products, and different users on different products. They built a logistic regression model to detect fraudulent reviews incorporating rating and textual features such as review title and body length, sentiment, cosine similarity between review and product texts, and others, and achieved an AUC of 78%. Similarly, Mukherjee et al. [61] leveraged cosine similarity across a user's given reviews and across a product's received reviews in addition to rating and temporal features in an unsupervised generative Bayesian model to automatically discern separating features of truthful and fraudulent reviewers (AUC = 0.86). Going beyond syntax, Sandulescu et al. [82] studied the problem of detecting singleton review spammers by comparing both review syntax and semantic similarity in pairwise reviews per business, and marked reviews with high similarity as fraudulent. Syntactic similarity was measured using part-of-speech tags and semantic similarity using word-to-word distances in the WordNet synonyms database. This approach achieved F1-score between 0.5 and 0.7 on Yelp and Trustpilot customer reviews data, and suggests that intelligent fraudsters often duplicate semantically similar messages by replacing some words between their fake reviews with synonymous or similar words in order to avoid generating blatant duplicates and be caught.

However, more complex review fraud exists, where fraudsters put considerably more effort than just duplicating review text in order to write sophisticated, deceptive reviews. To get ground truth deceptive reviews, Amazon Mechanical Turk (AMT) workers are frequently employed. The associated detection algorithms largely rely on text-processing and feature engineering for detecting such reviews. Ott et al. [70] collected 400 truthful and 400 positive-sentiment deceptive AMT-sourced reviews and trained Support Vector Machine (SVM) classifiers using a variety of feature sets, such as n-grams and LIWC features [72]. This achieved high 0.9 F1-score compared to human judges, who at best achieved a 0.7 F1-score. In a followup work [69], negative sentiment deceptive reviews were studied, with additional 400 negative reviews from AMT. Experiments showed that an SVM classifier trained on bigrams was again able to achieve strong performance in detecting such reviews, with a 0.86 F1-score. The authors additionally studied classifier performance when training on positive sentiment reviews and testing on negative sentiment reviews, and vice versa—results showed that such heterogeneity between training and testing data produced considerably worse F1-score (roughly a 0.1–0.2 reduction) than the homogeneous case, indicating different statistical patterns in positive and negative sentiment deceptive reviews versus truthful reviews.

While AMT generated reviews are common to use, they lack domain expertise. To address that, Li et al. [54] collected additional deceptive reviews from domain experts such as employees at target venues. The authors used n-gram features as in previous works and employ both Sparse Additive Generative Model (SAGE) and SVM classifiers to evaluate pairwise and three-class classification (truthful customer vs deceptive Turker vs deceptive employee) performance (65% accuracy). Their results showed that distinguishing truthful customers from deceptive employees is somewhat more difficult than from deceptive Turkers. Further, Li et al. [51] added sentiment, subjectivity, and pronoun usage features to the set. They then created a semi-supervised co-training algorithm that iteratively learns classifiers from review and reviewer features separately, and augments the training set in each successive iteration with the most agreed-upon and confidently scored reviews. This model achieves an F1-score of 0.63 on an Epinions review dataset.

A major aim of these systems is to aid humans in identifying fraud. Harris [33] focused on the usefulness of these models to human judges. Equipping human judges with these simple summary statistics of reviews improved their manual classification accuracy by up to 20% over the alternative (without), showing the effectiveness of augmented detection for humans at a cheaper computational cost.

6.1.2 **Graph-based detection**. These algorithms leverage the rating graph for identifying fake edges. Nodes in the graph represent users and products, and edge from u to p represents a review by user u to product p. Some algorithms also use features which may be available on nodes and/or edges.
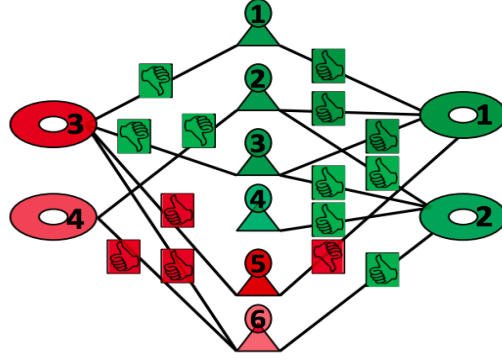
Fig. 12. Graph-based fake review detection algorithms are usually based on homophily, where good (green) users give positive "thumbs up" to other good products, while bad (red) users give negative "thumbs down" to them. The opposite is true for ratings given to bad products. Figure reprinted with permission from [5].

Belief propagation on the rating graph is one common way to identify fraud. Rayana et al. [79] used loopy belief propagation on the review graph network for identifying fake reviews, extending the idea of FraudEagle from Akoglu et al. [5]. The basic idea is presented in Figure 12. These algorithms take a signed network, i.e. a network where the edges are converted to be positive (thumbs up) and negative (thumbs down), and employ the notion of homophily which suggests that most honest users give genuine positive ratings to good products and negative ratings to bad products, and vice-versa for bad fraudulent users. This is expressed as a Markov Random Field, where the joint probability $P(\mathbf{y})$ of inferred labels Y_i is a product of entity i's prior beliefs $\phi_i(y_i)$ and its compatibility with labels of its neighbors j represented as $\gamma_{ij}^s(y_i, y_j)$, with the compatibility matrix s. Mathematically,

$$P(\mathbf{y}) = \frac{1}{Z} \ \Pi_{Y_i \in V} \phi_i(y_i) \ \Pi_{(Y_i, Y_j, s) \in E^{\pm}} \gamma_{ij}^s(y_i, y_j)$$

This is solved using loopy belief propagation, with prior-based initialization and transfer of beliefs across the network till convergence. Based on this idea, Rayana et al. [79] combines belief propagation with feature values of nodes and edges as well. This SpEagle algorithm is highly accurate in identifying fake reviews (and users) in three Yelp fake review datasets, with area under the ROC curve scores around 0.78 on average.

Several algorithms have been developed for jointly modeling user, review, and product information, with applications to fake review detection. Wang et al. [106] uses the review network to measure trustness of users $T(u)$, honesty of reviews $H(r)$, and reliability of stores $R(s)$, all of which lie between -1 and +1. For calculating $H(r)$ for a review r given to product p, the trustness of users S_r^+ and S_r^- who gave similar and different scores, respectively, to product p close in time to r. This is calculated as agreement score $A(r) = \sigma_{u_i \in S_r^+} T(u_i) - \sigma_{u_j \in S_r^-} T(u_j)$. Logistic functions are used to bound the scores in (-1, +1). The honest score $H(r)$ can then be used to identify fake reviews. The formulation is mutually interdependent as follows:

$$T(u) = \frac{2}{1 + e^{-H(r)}} - 1$$
$$H(r) = |R(s)| \ (\frac{2}{1 + e^{-A(r)}} - 1)$$
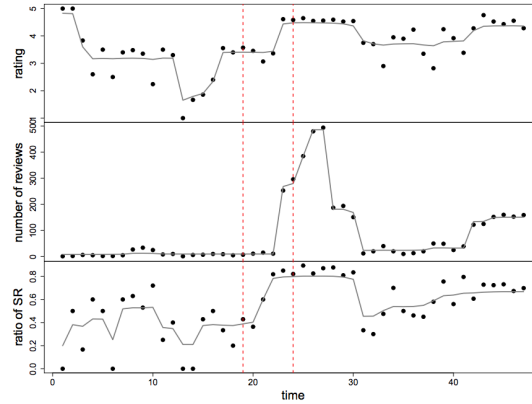$$R(s) = \frac{2}{1 + e^{-\Sigma_{(u, s) \in In(s)} T(u)}} - 1$$

Fig. 13. Figure showing suspicious rating time range (between red line) by showing bursty increase in rating (top), number of reviews (middle) and ratio of singleton review (bottom) in a coordinated manner. Figure reprinted with permission from [111].

The authors tested the algorithm to identify fake reviewers, which is a closely related problem, and get a precision of 49% on the 100 users with the smallest trustiness scores.

Closely related to the previous algorithm is Rev2 by Kumar et al. [48], which is also an iterative algorithm which calculates reviewer fairness, rating reliability and product goodness scores. The algorithm is based on the intuition that fraudulent review writers are typically unfair, in that they give unreliable rating scores to products that differ largely from the product's average score. Reliable reviewers give ratings that are close to the scores of other reliable reviewers. This algorithm incorporates user and product features by merging scores from a prior algorithm called Birdnest [36], and uses Bayesian priors for addressing cold start problems, i.e., judging the quality of users and products that only have a few ratings. This formulation is also interdependent, and the rating reliability is used to identify fake reviews. This algorithm achieves an AUC of over 0.85 for identifying fraudulent reviewers.

Several graph based algorithms have been developed to identify fraudulent nodes in review networks, using edge distributions [36, 84], dense block detection [16, 40], co-clustering [15], and more. This problem is closely related to identifying fake reviews, as the intuition is that by identifying fraudulent users, one can identify remove all their reviews and eliminate fake reviews. However, while these techniques may work for identifying bad users, these may not work well as-is in fake review detection because of two reasons: first, not all reviews by fraudulent users are necessarily fake [48] (for example, the user might aim to camouflage themselves by giving a few genuine reviews) and second, not all fake reviews are given by fraudulent users, which would hinder recall for fake review detection. Thus we do not focus on these algorithms in detail.

6.1.3 **Detection with temporal modeling**. This category includes algorithms that primarily using rating time information for identifying fake reviews. These approaches include a combination of feature engineering and modeling based on time series analysis, correlation, and co-clustering.

Ye et al. [115] considered the review data of each product as a stream of reviews, bucketed into temporal windows. For each temporal window, the authors consider a number of time series reflecting different properties: inter-rating time entropy, rating score entropy, average rating, review count, number of positive and negative reviews, and more. They created a LocalAR algorithm to identify anomalies in the time-series signal, which involves treating a single signal as the "lead" and using the rest of the signals as "supporting" ones. An abnormality

score at a timestep in the lead signal is defined as the residual between an empirically observed value and the forecasted value based on a local autoregressive (AR) model from previous timesteps. AR models generally take the following form:

$$X_t = \sum_{i=1}^{k} c_i X_{t-i} + \epsilon$$

The authors keep track of the distribution $D(S|T, P)$ of abnormality scores S over all timesteps T and products P, and define a threshold to flag abnormal timesteps, estimated as the percentage of expected anomalies using Cantelli's inequality. Upon finding anomalous points in the lead signal by this threshold, the algorithm turns to supporting signals for corroboration. For timesteps near only those flagged in the lead signal, the approach computes a local AR model on the supporting signals and calculates residual squared-error between estimates and empirically observed values at surrounding timesteps. If the residual error is above the abnormality threshold, the timestep is flagged as suspicious. These flags are integrated across the multiple signals using summary statistics like proportion of anomalies at a timestep. This LocalAR algorithm shows success on two case studies of bursty opinion spam on Flipkart data.

Along similar lines, Xie et al. [111] created CAPT-MDTS (Correlated Abnormal Patterns Detection in Multidimensional Time Series) based on burst detection in time-series. Specifically, the authors aim to find time periods during which a product is "under attack" by review fraudsters. The proposed approach involves detecting periods of time where there are correlated bursts in multiple time series reflecting average rating, review count, and proportion of singleton reviews. An illustration is given in Figure 13. The burst detection algorithm is built upon a variant of the longest common subsequence (LCS) problem which allows for two or more sequences (in this case, time series) to be considered "common" if they are approximately the same.

Furthermore, Li et al. [52] focused on detecting review fraud using both rating and spatiotemporal features in a supervised setting. The authors show that high average absolute rating deviation and high "average travel speed" (spatial distance between two subsequently reviewed venues divided by time between reviews) are suspicious, in addition to high distance between the registered location of the reviewer's account and the venue he/she is reviewing.

Overall, algorithms developed to identify opinion-based false information rely primarily on the information text, the entire user-review-product graph, and temporal sequence of reviews to identify individual and group of false reviews. Additional information, such as user properties, help as well. These algorithms have been developed and tested in a wide variety of platforms and datasets, and are efficient (high precision and AUC scores) in identifying fake reviews.

6.2 Detection of fact-based false information

In this part, we will look at the algorithms to detect hoaxes, fake news, and rumors in social media. These algorithms can be categorized into two major categories: feature engineering based and propagation based. Similar to opinion-based feature engineering methods, here features are created from their textual properties, their relation to other existing information, the properties of the users interacting with this information (e.g., the creator), and propagation dependent features (e.g., number of users that reshare a tweet). Feature-based algorithms have been used to identify various different types of malicious users and activities, such as identifying bots [97], trolls [20], vandals [49], sockpuppets [47], and many more.

Fact-based false information propagates through social networks, as opposed to opinion-based false information. Thus, propagation based algorithms model how true information propagates in these networks and anomalies of these models are predicted as false information. Some algorithms also create separate models for true and false

information propagation. Alternatively, propagation structures and information can be fed into machine learning models for prediction as well.

We will first discuss feature based algorithms (Section 6.2.1) and then explain propagation based models in Section 6.2.2.

6.2.1   *Feature-based detection*. **Text-based features:**

Text-analysis is core to identifying misinformation as the information being conveyed is primarily textual content. Similarly to opinion-based textual detection methods, research papers in this category are predominantly feature-based, where features can broadly be categorized as either stylometric (e.g., number of characters in a word), psycholinguistic (e.g., LIWC [72]), or complexity-oriented (e.g., readability indices).

One of the first studies on identifying rumors on Twitter was done by Qazninian et al. [77]. They collected manual annotations for over 10,000 tweets, and developed three categories of features to identify the false tweets—primarily based on content (unigram, bigrams, and part-of-speech), but also used user information (whether user has previously posted false information), and Twitter-specific information (hashtags and URLs). These features were converted into their log-likelihood ratio of being from the true or false class based on their distribution in the training data, and the combined score was used for classification. This model achieved a mean average precision score of 95%, indicating near-perfect classification. Individually, content features performed the best, followed by network features, and lastly hashtag and URL based Twitter features. Content based features were also the best performing features in Gupta et al. [32], which focused on fake tweet detection as well.

A recent study on false news detection by Perez-Rosas et al. [74] shows the changing effectiveness of text-based features. They collected a large dataset of false information generated by Amazon Mechanical Turk (AMT) workers and another dataset of celebrity fake news from gossip websites. They used a huge set of text based features for classification, consisting of n-grams, punctuations, LIWC, readability, and syntax features. Since the experiments were conducted on a balanced dataset, the baseline accuracy is 50%, and the combined set of features achieves an average accuracy of 74%. Unsurprisingly, cross-domain analysis, i.e., training model on one dataset and testing on another, dropped the accuracy to 61%. Within the same dataset, but training and testing on different domains of fake news (e.g., technology, education, etc.) the performance lies between 75% and 85%, indicating that knowledge can be transferred from one domain to another with reasonable performance.

Similarly, in their study to identify fake from real news from text, Horne and Adali [37] achieved accuracies between 71%–78%. Identifying fake news from satire was more difficult than identifying fake news from real news, indicating that the former two are written quite similarly. In their experiments, they observe that the news body text is less informative than the news title text in distinguishing real from fake information. This is an important finding and opens avenues of future research to quantify the dissonance between the title and body of an information piece, and using it as an indicator of false information.

The above studies show an alarming trend. In the earlier research papers, content-based detection performed well, but more recent research has shown that content alone is not good enough. This suggests a trend that malicious users and content creators are adapting and becoming more aware of how to create more genuine-looking false information which fools automated classifiers.

User, network, and metadata based detection:
These detection models derive features from several different aspects of data which we describe below. We start by looking at the features developed to identify hoaxes in Wikipedia. Kumar et al. [50] developed four different categories of features to identify hoaxes: (i) *appearance features*, that measures the length of the article, number of references, ratio of text to Wikipedia markup and so on, (ii) *network features*, that measures the relation between the references of the article in the Wikipedia hyperlink network, (iii) *support features*, quantifying the number of previous occurrences of the article title in other articles, and the time since its first occurrence to the time the article is created, and (iv) *creator features*, i.e., properties of article creator in terms of its previous edit
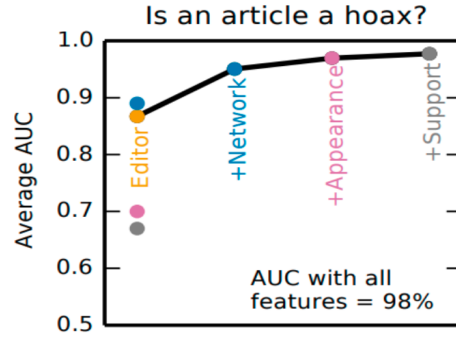
Fig. 14. Plot showing accuracy of hoax detection using different feature sets. Reprinted with permission from [50].
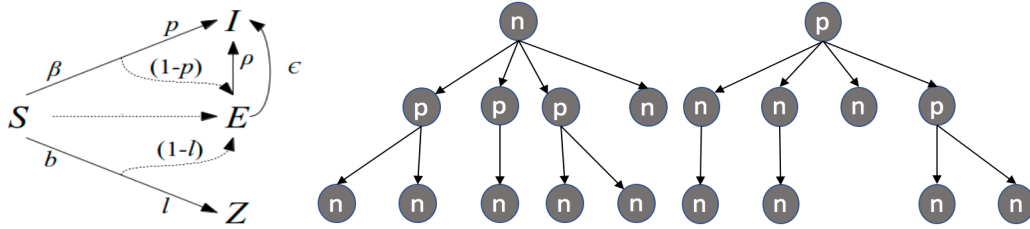

Fig. 15. (a) SEIZ model of information propagation. Reprinted with permission from [41]. (b–c) Common information propagation structure or motifs for (b) false information, and (c) real information. Adapted from [110].

count, time since registration, and so on. Figure 14 shows the performance of the features using a random forest classifier. Individually, creator and network features perform equally well (AUC = 0.90) Further, the performance can be improved significantly when other sets of features are incrementally added. The combination of all four categories of features gives an AUC of 0.98, indicating a near perfect classifier. This shows that in order to identify false information, one needs to look beyond its appearance and dig deeper into who created the piece of false information and how it relates to existing information—simply looking at its appearance is not as effective.

In the domain of fake Twitter images during disasters, Gupta et al. [32] used user features (number of friends, account age, etc.), tweet features (number of words, sentiment, POS, LIWC, etc.), and metadata features (number of hashtags, mentions, URLs, retweets). The dataset had 11,534 tweets, half of which were fake and other half were real, and decision trees were used for classification. User features alone had close to random accuracy of about 53%, while tweet features alone got near perfect accuracy of 97.7%, indicating that in the propagation of false information, what the tweet says is more important than the user who tweets it.

Thus, feature engineering has proven successful in identifying fake from true rumors and hoaxes, primarily using features derived from the text, user, network, and other metadata. These algorithms typically have performance numbers in high 80s and 90s, showing that they are effective and practically useful.

6.2.2 **Detection using propagation modeling**. The spread of information across social media adds another dimension to use for identification of true information from false information. A common way of using propagation information is to create models that broadly serve two purposes: to closely model the spread of (false) information, and to find ways to prevent its spread.

Creating true information propagation models:
Acemoglu et al. [3] presented one of the first models to simulate propagation of information in social media. They considered information spread as exchange of belief about information (e.g., supporting a political candidate), and theoretically find the cases in which false information survives in the network. Nodes in the network are considered to be either normal or forceful. When two normal nodes interact, each of them adopts an average of their existing beliefs. But interactions with forceful nodes only change the belief of the normal node, while forceful node only slightly updates its beliefs. With this interaction model, simulations showed that belief about false information dies out when the social network is well connected and normal nodes interact with each other a lot. On the other hand, *echo chambers* are formed and false information prevails when there are several forceful nodes who update their own belief from the beliefs of nodes they previously influenced. This model suggests increasing the number of normal nodes and increasing their connectivity with each other may be a way to reduce false information propagation.

More recently, Jin et al. [41] created a false information propagation model which they called SEIZ. Similar to the standard SIS model, each node in SEIZ model lies in one of four states—susceptible (S), exposed (E), infected (I), and skeptical (Z), as shown in Figure 15(a). Initially, nodes are susceptible (state S). When they receive information, they can transition to states I or Z with probabilities β and b, respectively. These transitions may only be successful with some probabilities p and b, respectively, otherwise the nodes transition to state E. The following four equations explain the transitions according to this SEIZ model:

$$\frac{d[S]}{dt} = -\beta S \frac{I}{N} - b S \frac{Z}{N}$$
$$\frac{d[E]}{dt} = (1-p)\beta S \frac{I}{N} + (1-l)b S \frac{Z}{N} - \rho E \frac{I}{N} - \epsilon E$$
$$\frac{d[I]}{dt} = p\beta S \frac{I}{N} + \rho E \frac{I}{N} + \epsilon E$$
$$\frac{d[Z]}{dt} = lb S \frac{Z}{N}$$

The model parameters are learned by training on real true and false information propagation data, and these are used for prediction of false information. A metric measuring the rate of users entering state E to leaving it is predictive of false information—a high value indicates true information while a low score indicates a rumor, as shown by their case study of eight rumors.

Incorporating propagation information in machine learning models:
Apart from model creation, propagation information and propagation structure can both augment existing machine learning frameworks. For example, Yang et al. [112] added propagation features such as number of replies and number of retweets to the standard set of features used for their classification tasks, such as content, client (device type), user, and location, to identify false information spread on the Sina Weibo network. Using this feature set with an SVM classifier gave an average accuracy of 78%.

A more structured way of using propagation information was created by Wu et al. [110], who also studied 11,466 rumors on Sina Weibo by using propagation information. Each thread of information was represented as a tree, with the original message as the root and its replies as children, and so on. This is shown in Figure 15 (b) and (c) for false and true information, respectively. Popular nodes, i.e., ones with at least 1,000 followers, are denoted as p and others as n, to understand if popular nodes boost false information. The authors observe that false information is usually started by a normal user, then reported and supported by some opinion leaders, and then finally reshared by a large number of normal users (Figure 15(b)). On the other hand, true information is posted by opinion leaders and then reposted directly by normal users (Figure 15(c)). With features representing the propagation structure, user features, average sentiment, doubt, surprise and emotion, an SVM classifier achieved

91% accuracy. Further, early detection of false information achieved 72% accuracy even without any propagation information, and 88% with propagation information of its first 24 hours.

Mitigation by modeling true and false information:
Previously we described propagation models spreading false information. Here, we consider models that model the spread of both true and false information simultaneously, where success is measured as the number of users saved from accepting false information. The aim of these models is to create mitigation strategies to reduce and prevent the spread of false information. Several such models have been developed. Tripathy et al. [103] created two models in which true information (anti-rumor) is spread after false information(rumor) starts to spread, based on real data. In one model, truth is spread n days after falsehood to simulate real-world observed time-lag, and in the second model, truth is spread by some special nodes (e.g., official accounts) as soon as they receive false information. They conducted experiments with Twitter and simulated networks, and find that there is a super-linear relationship between the lifetime of a rumor and delay of its detection. When the special nodes detect and spread anti-rumors, it reduces the average lifetime of rumors by over 60%.

Similarly, Budak et al. [18] presented an independent cascade model called Multi-Campaign Independence Cascade Model (MCICM). The model contains a rumor campaign and a true information 'limiting' campaign spreading through the network. Each node, whenever infected with true or false information, spreads its belief to its neighbors, which accept the information with some probability. Their algorithm learns the model parameters, even with missing data. Their simulations show that a 30% increase in delay in starting the true information spread reduces its reach by 80% of the population. More recently, Nguyen et al. [64] created another model with both linear threshold and independent cascade, and found that when true information can only be spread by a few nodes, it is most effective to do it via highly influential nodes in large communities. When more nodes can be selected, influential nodes in smaller communities are more effective in preventing the spread of false information. This method is 16–41% better than other methods. Related models have also been developed [119].

Thus, several propagation models have been developed that capture the spread of true and false information on social media. These models are used independently or in conjunction with other machine learning algorithms to identify false information. These algorithms are effective in detecting spread of rumors, and their simulations suggest rumor mitigation strategies.

Overall, several categories of false information detection algorithms have been developed for opinion-based and fact-based false information. The common category of algorithm is feature-based, which converts observations into feature vectors, derived from text, user, network, and metadata. Graph-based algorithms have primarily been developed for opinion-based false information (e.g., fake reviews), and identify dense block of users or information, potentially also occurring in bursty short time period. Temporal modeling algorithms use time-series analysis, and co-clustering on one or more of time evolving properties of information (e.g., number of ratings per day) to identify opinion-based false information as well. Fact-based information that spreads (e.g., rumors) is also detected by creating true and false information propagation models. All these types of algorithms perform well in their respective datasets to identify false information, and usually achieve a high accuracy, precision, or AUC score in the 80s or 90s.

7   DISCUSSIONS AND OPEN CHALLENGES

In this survey, we took a comprehensive view on mechanisms, rationale, impact, characteristics, and detection of three types of false information: fake reviews, hoaxes, and fake news.

Several algorithms have been developed for detection in different domains. However, they are not directly comparable to each other due to the lack of large-scale publicly available datasets of false information, spanning fake reviews, hoaxes, and social media rumors. This prevents a benchmark comparison between different categories of algorithms. Such datasets are needed to understand the advantages and disadvantages of different

algorithms, and collectively improve the state of the art. Some recent datasets, such as from Buzzfeed [92], LIAR [107], and CREDBANK [59], and FakeNewsNet [89, 90], have been created but standardized comparison of existing algorithms on these datasets has not been conducted.

The next generation of false information will be fueled by the advancements in machine learning. Recent research has shown that machine learning models can be built to create genuine looking text [114], audio [63], images and videos [99]. With further development of such techniques, it will become increasingly difficult for readers to identify false from true information. Techniques to separate the two using standard signals, like text, user information, group-level interaction, time, and more will need newer reforms for combating this new wave.

There are several open avenues of research, spanning areas of machine learning, natural language processing, signal processing, information retrieval, big data analysis, and computer vision for studying, characterizing, detecting and preventing false information on the web and social media:

Semantic dissonance detection: Some smartly created false information pieces cite references to look credible, but the reference may not reflect what the information piece says. Often, the summary of some information (e.g. the headline of a news information piece) may not convey the same message as the main information itself. These tactics leverage human laziness and aversion to fact-checking. Natural language processing algorithms for semantic dissonance estimation can help identify such deceptive sources.

Fact-checking from knowledge bases: Fact-based false information can be checked by matching it against a knowledge base of complete information. This direction poses numerous challenges. The first is successfully creating and maintaining this knowledge-base, and ascertaining data quality. Next, natural language understanding and information extraction techniques must be developed to automatically extract information from free-form natural text. Finally, we would require capable information matching algorithms in order to check if the extracted information matched with the existing information in the knowledge base.

Fact-checking using crowdsourcing: Readers express different emotions, such as skepticism, when interacting with false information as compared to true information [105]. They may also "report" such posts. Manual fact-checking of all stories is not feasible and such stories can be created to bypass existing detection filters, which is where crowdsourced signals can be useful. These algorithms can help in early detection of fake news [44] and resource allocation of fact-checkers.

Multimedia false information detection: As demonstrated by recent research, fabricated and manipulated audio [63], images and videos [99] can be developed using learning technologies. Research topics in these directions include developing signal processing, computer vision, and data analysis techniques to identify signature characteristics of fabricated or manipulated multimedia, and developing machine learning algorithms for their detection.

Bridging echo chambers: The formation of social media echo chambers fuels the presence and spread of false information. One strategy to combat false information is to bridge conflicting echo chambers, so that opposing information can be exchanged and considered. Data-driven models of effective means of bridging these echo chambers/filter bubbles are needed. At the same time, further research is required in order to effectively present opposing beliefs to readers in order to reduce polarization.

Adversarial creation of false information: Malicious users that try to create and spread false information are actively involved in the process. They can adapt their future behavior based on the counter-measures that are being taken to detect and prevent their current behaviors. Therefore, research in the direction of dealing with adaptive adversaries is promising in mitigating the impact of false information.

Mitigation of false information: Reducing the damage of false information is an essential direction that is open for research. Very recent research has shown that educating people against possible manipulation strategies used in false information is effective in improving human detection skills [104]. Further research in finding effective educational strategies to "vaccinate" people against believing false information, and how to scale these strategies to millions of users that use social platforms is necessary.

ACKNOWLEDGEMENT

REFERENCES

[1] Planet money: Episode 739: Finding the fake-news king. http://www.npr.org/templates/transcript/transcript.php?storyId=504155809.

[2] Why we study digital misinformation. http://cnets.indiana.edu/blog/2016/11/28/why-we-study-digital-misinformation/.

[3] Daron Acemoglu, Asuman Ozdaglar, and Ali ParandehGheibi. Spread of (mis) information in social networks. *Games and Economic Behavior*, 70(2):194–227, 2010.

[4] Scott F Aikin. *Poe's Law, Group Polarization, and the Epistemology of Online Religious Discourse.* 2009.

[5] Leman Akoglu, Rishi Chandy, and Christos Faloutsos. Opinion fraud detection in online reviews by network effects. *Proceedings of the 7th International AAAI Conference on Web and Social Media*, 2013.

[6] Leman Akoglu, Mary McGlohon, and Christos Faloutsos. Oddball: Spotting anomalies in weighted graphs. In *Advances in Knowledge Discovery and Data Mining*. Springer, 2010.

[7] Jonathan Albright. Data is the real post-truth, so here's the truth about post-election2016 propaganda. https://medium.com/@d1gi/data-is-the-real-post-truth-so-heres-the-truth-about-post-\election2016-propaganda-2bff5ae1dd7.

[8] Jonathan Albright. The election2016 micro-propaganda machine. https://medium.com/@d1gi/the-election2016-micro-propaganda-machine-383449cc1fba.

[9] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. Technical report, National Bureau of Economic Research, 2017.

[10] Solomon E Asch and H Guetzkow. Effects of group pressure upon the modification and distortion of judgments. *Groups, leadership, and men*, pages 222–236, 1951.

[11] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, 2011.

[12] Tim Berners-Lee, Robert Cailliau, Jean-François Groff, and Bernd Pollermann. World-wide web: The information universe. *Internet Research*, 20(4):461–471, 2010.

[13] Alessandro Bessi and Emilio Ferrara. Social bots distort the 2016 us presidential election online discussion. *First Monday*, 2016.

[14] Alex Beutel. User behavior modeling with large-scale graph analysis. 2016.

[15] Alex Beutel, Kenton Murray, Christos Faloutsos, and Alexander J Smola. CoBaFi: collaborative bayesian filtering. In *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 2014.

[16] Alex Beutel, Wanhong Xu, Venkatesan Guruswami, Christopher Palow, and Christos Faloutsos. Copycatch: stopping group attacks by spotting lockstep behavior in social networks. In *Proceedings of the 22rd International Conference on World Wide Web*. ACM, 2013.

[17] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8, 2011.

[18] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. Limiting the spread of misinformation in social networks. In *Proceedings of the 20th International Conference on World Wide Web*. ACM, 2011.

[19] Justin Cheng, Lada Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. Can cascades be predicted? In *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 2014.

[20] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Antisocial behavior in online discussion communities. In *Proceedings of the 9th International AAAI Conference on Web and Social Media*, 2015.

[21] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*, 2016.

[22] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559, 2016.

[23] Benjamin Doerr, Mahmoud Fouz, and Tobias Friedrich. Why rumors spread so quickly in social networks. *Communications of the ACM*, 55(6):70–75, 2012.

[24] Don Fallis. A conceptual analysis of disinformation. *Proceedings of iConference*, 2009.

[25] Don Fallis. A functional analysis of disinformation. *Proceedings of iConference*, 2014.

[26] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *Communications of the ACM*, 59(7):96–104, 2016.

[27] Marc Fisher, John Woodrow Cox, and Peter Hermann. Pizzagate: From rumor, to hashtag, to gunfire in dc. *Washington Post*, 2016.

[28] DJ Flynn, Brendan Nyhan, and Jason Reifler. The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology*, 38(S1):127–150, 2017.

[29] Michelle C Forelle, Philip N Howard, Andrés Monroy-Hernández, and Saiph Savage. Political bots and the manipulation of public opinion in venezuela. *SSRN Electronic Journal*, 2015.

[30] Adrien Friggeri, Lada A Adamic, Dean Eckles, and Justin Cheng. Rumor cascades. In *Proceedings of the 8th International AAAI Conference on Web and Social Media*, 2014.

[31] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Balancing opposing views to reduce controversy. *Proceedings of the 10th ACM International Conference on Web Search and Data Mining*, 2017.

[32] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 2013.

[33] Christopher Harris. Detecting deceptive opinion spam using human computation. In *Proceedings of the Workshops at AAAI on Artificial Intelligence*, 2012.

[34] Alfred Hermida. Twittering the news: The emergence of ambient journalism. *Journalism practice*, 4(3):297–308, 2010.

[35] Peter Hernon. Disinformation and misinformation through the internet: Findings of an exploratory study. *Government Information Quarterly*, 12(2):133–139, 1995.

[36] Bryan Hooi, Neil Shah, Alex Beutel, Stephan Günnemann, Leman Akoglu, Mohit Kumar, Disha Makhija, and Christos Faloutsos. Birdnest: Bayesian inference for ratings-fraud detection. In *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, 2016.

[37] Benjamin D. Horne and Sibel 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 *The 2nd International Workshop on News and Public Opinion*, 2017.

[38] Philip N Howard and Bence Kollanyi. Bots, #strongerin, and #brexit: Computational propaganda during the uk-eu referendum. *SSRN Electronic Journal*, 2016.

[39] Lee Howell et al. Digital wildfires in a hyperconnected world. *WEF Report*, 3:15–94, 2013.

[40] Meng Jiang, Peng Cui, Alex Beutel, Christos Faloutsos, and Shiqiang Yang. Inferring strange behavior from connectivity pattern in social networks. In *Advances in Knowledge Discovery and Data Mining*. Springer, 2014.

[41] Fang Jin, Edward Dougherty, Parang Saraf, Yang Cao, and Naren Ramakrishnan. Epidemiological modeling of news and rumors on twitter. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis*. ACM, 2013.

[42] Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. News verification by exploiting conflicting social viewpoints in microblogs. In *AAAI*, pages 2972–2978, 2016.

[43] Nitin Jindal and Bing Liu. Opinion spam and analysis. In *Proceedings of the 1st ACM International Conference on Web Search and Data Mining*. ACM, 2008.

[44] Jooyeon Kim, Behzad Tabibian, Alice Oh, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 324–332. ACM, 2018.

[45] David Krech and Richard S Crutchfield. *Perceiving the World*. McGraw-Hill, 1948.

[46] Mohit Kumar and Disha Makhija. Reviews and ratings fraud detection in e-commerce, 2015.

[47] Srijan Kumar, Justin Cheng, Jure Leskovec, and VS Subrahmanian. An army of me: Sockpuppets in online discussion communities. In *Proceedings of the 26th International Conference on World Wide Web*, 2017.

[48] Srijan Kumar, Bryan Hooi, Disha Makhija, Mohit Kumar, Christos Faloutsos, and VS Subrahamanian. Rev2: Fraudulent user prediction in rating platforms. *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, 2018.

[49] Srijan Kumar, Francesca Spezzano, and VS Subrahmanian. Vews: A wikipedia vandal early warning system. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015.

[50] Srijan Kumar, Robert West, and Jure Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th International Conference on World Wide Web*, 2016.

[51] Fangtao Li, Minlie Huang, Yi Yang, and Xiaoyan Zhu. Learning to identify review spam. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2011.

[52] Huayi Li, Zhiyuan Chen, Arjun Mukherjee, Bing Liu, and Jidong Shao. Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns. In *Proceedings of the 9th International AAAI Conference on Web and Social Media*, 2015.

[53] Huayi Li, Geli Fei, Shuai Wang, Bing Liu, Weixiang Shao, Arjun Mukherjee, and Jidong Shao. Bimodal distribution and co-bursting in review spam detection. In *Proceedings of the 26th International Conference on World Wide Web*, 2017.

[54] Jiwei Li, Myle Ott, Claire Cardie, and Eduard Hovy. Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014.

[55] Yuming Lin, Tao Zhu, Xiaoling Wang, Jingwei Zhang, and Aoying Zhou. Towards online review spam detection. In *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 2014.

[56] Michael Luca and Georgios Zervas. Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science*, 62(12):3412–3427, 2016.

[57] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter under crisis: Can we trust what we RT? *Proceedings of the First Workshop on Social Media Analytics*, 2010.

[58] Amanda J Minnich, Nikan Chavoshi, Abdullah Mueen, Shuang Luan, and Michalis Faloutsos. Trueview: Harnessing the power of multiple review sites. In *Proceedings of the 24th International Conference on World Wide Web*, 2015.

[59] Tanushree Mitra and Eric Gilbert. Credbank: A large-scale social media corpus with associated credibility annotations. In *Proceedings of the 9th International AAAI Conference on Web and Social Media*, 2015.

[60] Tanushree Mitra, Graham P Wright, and Eric Gilbert. A parsimonious language model of social media credibility across disparate events. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 2017.

[61] Arjun Mukherjee, Bing Liu, and Natalie Glance. Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st International Conference on World Wide Web*. ACM, 2012.

[62] Arjun Mukherjee, Vivek Venkataraman, Bing Liu, and Natalie S Glance. What yelp fake review filter might be doing? In *Proceedings of the 7th International AAAI Conference on Web and Social Media*, 2013.

[63] Dibya Mukhopadhyay, Maliheh Shirvanian, and Nitesh Saxena. All your voices are belong to us: Stealing voices to fool humans and machines. In *Proceedings of European Symposium on Research in Computer Security*. Springer, 2015.

[64] Nam P Nguyen, Guanhua Yan, My T Thai, and Stephan Eidenbenz. Containment of misinformation spread in online social networks. In *Proceedings of the 4th Annual ACM Web Science Conference*. ACM, 2012.

[65] Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175, 1998.

[66] A Conrad Nied, Leo Stewart, Emma Spiro, and Kate Starbird. Alternative narratives of crisis events: Communities and social botnets engaged on social media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing companion*. ACM, 2017.

[67] Dimitar Nikolov, Diego FM Oliveira, Alessandro Flammini, and Filippo Menczer. Measuring online social bubbles. *PeerJ Computer Science*, 1:e38, 2015.

[68] Brendan Nyhan and Jason Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330, 2010.

[69] Myle Ott, Claire Cardie, and Jeffrey T Hancock. Negative deceptive opinion spam. In *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013.

[70] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.

[71] Shashank Pandit, Duen Horng Chau, Samuel Wang, and Christos Faloutsos. Netprobe: a fast and scalable system for fraud detection in online auction networks. In *Proceedings of the 16th International Conference on World Wide Web*. ACM, 2007.

[72] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

[73] Gordon Pennycook and David Rand. *Who Falls for Fake News? The Roles of Analytic Thinking, Motivated Reasoning, Political Ideology, and Bullshit Receptivity*, 2017 (accessed August 27, 2017).

[74] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*, 2017.

[75] Andrew Perrin. Social media usage. *Pew Research Center*, 2015.

[76] Peter Pomerantsev and Michael Weiss. *The menace of unreality: How the Kremlin weaponizes information, culture and money*. Institute of Modern Russia New York, 2014.

[77] Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 2011.

[78] Walter Quattrociocchi, Antonio Scala, and Cass R Sunstein. *Echo chambers on facebook.* 2016.

[79] Shebuti Rayana and Leman Akoglu. Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015.

[80] Victoria L Rubin, Niall J Conroy, Yimin Chen, and Sarah Cornwell. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the 14th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016.

[81] Natali Ruchansky, Sungyong Seo, and Yan Liu. Csi: A hybrid deep model for fake news. *arXiv preprint arXiv:1703.06959*, 2017.

[82] Vlad Sandulescu and Martin Ester. Detecting singleton review spammers using semantic similarity. In *Proceedings of the 24th international conference on World Wide Web*. ACM, 2015.

[83] Neil Shah. FLOCK: Combating astroturfing on livestreaming platforms. In *Proceedings of the 26th International Conference on World Wide Web*, 2017.

[84] Neil Shah, Alex Beutel, Bryan Hooi, Leman Akoglu, Stephan Gunnemann, Disha Makhija, Mohit Kumar, and Christos Faloutsos. Edgecentric: Anomaly detection in edge-attributed networks. In *Proceedings of the 2016 IEEE International Conference on Data Mining Workshops*, 2016.

[85] Neil Shah, Hemank Lamba, Alex Beutel, and Christos Faloutsos. The many faces of link fraud. In *Proceedings of the 2017 IEEE International Conference on Data Mining*, 2017.

[86] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th International Conference Companion on World Wide Web*, 2016.

[87] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Alessandro Flammini, and Filippo Menczer. The spread of fake news by social bots. *arXiv preprint arXiv:1707.07592*, 2017.

[88] Elisa Shearer and Jeffrey Gottfried. News use across social media platforms 2017. *Pew Research Center*, September 2017.

[89] Kai Shu, Suhang Wang, and Huan Liu. Exploiting tri-relationship for fake news detection. *arXiv preprint arXiv:1712.07709*, 2017.

[90] Kai Shu, Suhang Wang, Amy Sliva, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *arXiv preprint arXiv:1708.01967*, 2017.

[91] Craig Silverman. Lies, damn lies, and viral content. how news websites spread (and debunk) online rumors, unverified claims, and misinformation. *Tow Center for Digital Journalism*, 168, 2015.

[92] Craig Silverman. This analysis shows how viral fake election news stories outperformed real news on facebook. *Buzzfeed News*, 16, 2016.

[93] Brian Skyrms. *Signals: Evolution, learning, and information.* Oxford University Press, 2010.

[94] Alexander Smith and Vladimit Banic. Fake news: How a partying macedonian teen earns thousands publishing lies. *NBC News*, 2016.

[95] Kate Starbird. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In *Proceedings of the 12th International AAAI Conference on Web and Social Media*, 2017.

[96] Kate Starbird, Jim Maddock, Mania Orand, Peg Achterman, and Robert M Mason. Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing. *iConference 2014 Proceedings*, 2014.

[97] VS Subrahmanian, Amos Azaria, Skylar Durst, Vadim Kagan, Aram Galstyan, Kristina Lerman, Linhong Zhu, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. The darpa twitter bot challenge. *IEEE Computer*, 49(6):38–46, 2016.

[98] VS Subrahmanian and Srijan Kumar. Predicting human behavior: The next frontiers. *Science*, 355(6324):489–489, 2017.

[99] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *Journal of ACM Transactions on Graphics*, 36(4):95, 2017.

[100] Eugenio Tacchini, Gabriele Ballarin, Marco L Della Vedova, Stefano Moret, and Luca de Alfaro. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506*, 2017.

[101] Jeffrey E Thomas. Statements of fact, statements of opinion, and the first amendment. *California Law Review*, 74(3):1001–1056, 1986.

[102] David Trilling. *Seen a fake news story recently? You're more likely to believe it next time*, 2017 (accessed August 27, 2017).

[103] Rudra M Tripathy, Amitabha Bagchi, and Sameep Mehta. A study of rumor control strategies on social networks. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. ACM, 2010.

[104] Sander van der Linden, Anthony Leiserowitz, Seth Rosenthal, and Edward Maibach. Inoculating the public against misinformation about climate change. *Global Challenges*, 1(2), 2017.

[105] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[106] Guan Wang, Sihong Xie, Bing Liu, and S Yu Philip. Review graph based online store review spammer detection. In *Proceedings of the 2011 IEEE International Conference on Data Mining*. IEEE, 2011.

[107] William Yang Wang. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*, 2017.

[108] Andrew Ward, L Ross, E Reed, E Turiel, and T Brown. Naive realism in everyday life: Implications for social conflict and misunderstanding. *Values and knowledge*, pages 103–135, 1997.

[109] Tim Weninger, Thomas James Johnston, and Maria Glenski. Random voting effects in social-digital spaces: A case study of reddit post submissions. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. ACM, 2015.

[110] Ke Wu, Song Yang, and Kenny Q Zhu. False rumors detection on sina weibo by propagation structures. In *Proceedings of the 31st IEEE International Conference on Data Engineering*. IEEE, 2015.

[111] Sihong Xie, Guan Wang, Shuyang Lin, and Philip S Yu. Review spam detection via temporal pattern discovery. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 823–831. ACM, 2012.

[112] Fan Yang, Yang Liu, Xiaohui Yu, and Min Yang. Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*. ACM, 2012.

[113] Zhi Yang, Christo Wilson, Xiao Wang, Tingting Gao, Ben Y Zhao, and Yafei Dai. Uncovering social network sybils in the wild. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(1):2, 2014.

[114] Yuanshun Yao, Bimal Viswanath, Jenna Cryan, Haitao Zheng, and Ben Zhao. Automated crowdturfing attacks and defenses in online review systems. In *Proceedings of the ACM Conference on Computer and Communications Security*, 2017.

[115] Junting Ye, Santhosh Kumar, and Leman Akoglu. Temporal opinion spam detection by multivariate indicative signals. In *Proceedings of the 10th International AAAI Conference on Web and Social Media*, 2016.

[116] Robert B Zajonc. Attitudinal effects of mere exposure. *Journal of personality and social psychology*, 9(2p2):1, 1968.

[117] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. The web centipede: Understanding how web communities influence each other through the lens of mainstream and alternative news sources. *Proceedings of the 2017 ACM Internet Measurement Conference*, 2017.

[118] Li Zeng, Kate Starbird, and Emma S Spiro. Rumors at the speed of light? modeling the rate of rumor transmission during crisis. In *Proceedings of the 49th Hawaii International Conference on System Sciences*. IEEE, 2016.

[119] Laijun Zhao, Hongxin Cui, Xiaoyan Qiu, Xiaoli Wang, and Jiajia Wang. Sir rumor spreading model in the new media age. *Physica A: Statistical Mechanics and its Applications*, 392(4):995–1003, 2013.

[120] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. Detection and resolution of rumours in social media: A survey. *arXiv preprint arXiv:1704.00656*, 2017.

[121] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989, 2016.