Combating Fake News: A Survey on Identification and Mitigation Techniques

KARISHMA SHARMA, FENG QIAN, HE JIANG, and NATALI RUCHANSKY,
University of Southern California
MING ZHANG, Peking University
YAN LIU, University of Southern California

The proliferation of fake news on social media has opened up new directions of research for timely identification and containment of fake news and mitigation of its widespread impact on public opinion. While much of the earlier research was focused on identification of fake news based on its contents or by exploiting users' engagements with the news on social media, there has been a rising interest in proactive intervention strategies to counter the spread of misinformation and its impact on society. In this survey, we describe the modern-day problem of fake news and, in particular, highlight the technical challenges associated with it. We discuss existing methods and techniques applicable to both identification and mitigation, with a focus on the significant advances in each method and their advantages and limitations. In addition, research has often been limited by the quality of existing datasets and their specific application contexts. To alleviate this problem, we comprehensively compile and summarize characteristic features of available datasets. Furthermore, we outline new directions of research to facilitate future development of effective and interdisciplinary solutions.

CCS Concepts: • **Information systems** → *Social networking sites*; *Data mining*; • **Computing methodologies** → *Machine learning*;

Additional Key Words and Phrases: AI, fake news detection, rumor detection, misinformation

ACM Reference format:
Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. 2019. Combating Fake News: A Survey on Identification and Mitigation Techniques. *ACM Trans. Intell. Syst. Technol.* 10, 3, Article 21 (April 2019), 42 pages.
https://doi.org/10.1145/3305260

Feng Qian was a visiting student at University of Southern California.

This work is supported in part by the NSF Research Grant IIS-1619458 and IIS-1254206 as well as the National Natural Science Foundation of China NSFC Grant (NSFC Grant Nos.61772039, 91646202 and 61472006). The views and conclusions are those of the authors and should not be interpreted as representing the social policies of the funding agency, or the U.S. Government.

Authors' addresses: K. Sharma, F. Qian, H. Jiang, N. Ruchansky, and Y. Liu, University of Southern California, 3650 McClintock Ave, Los Angeles, California, USA, 900089; emails: krsharma@usc.edu, nickqian@pku.edu.cn, jian567@usc.edu, natalir@bu.edu, yanliu.cs@usc.edu; M. Zhang, Peking University, Beijing, China, 100871; email: mzhang@net.pku.edu.cn.

2157-6904/2019/04-ART21 $15.00
https://doi.org/10.1145/3305260

1 INTRODUCTION

In recent years, the topic of *fake news* has experienced a resurgence of interest in society. The increased attention stems largely from growing concerns around the widespread impact of fake news on public opinion and events. In January 2017, a spokesman for the German government stated that they "are dealing with a phenomenon of a dimension that [they] have not seen before," referring to the proliferation of fake news on social media.[1] Although social media has increased the ease with which real-time information disseminates, its popularity has exacerbated the problem of fake news by expediting the speed and scope at which false information can be spread. Fuller et al. [32] noted that with the massive growth of online communication, the potential for people to deceive through computer-mediated communication has also grown, and such deception can have disastrous and far-reaching results on many areas of our lives, including financial markets [14, 53] and political events [2, 4]. For instance, Carvalho et al. [14] noted that a false report of bankruptcy of a United Airlines parent company in 2008 caused the stock price to drop by as much as 76% in a matter of minutes; although the stock rebounded after the news was identified as false, it closed 11.2% lower than the previous day and the negative effect persisted for 6 more days. In terms of political events, an analysis of the US presidential election in 2016 by Allcott and Gentzkow [2] revealed that fake news was widely shared during the 3 months prior to the election with 30 million total Facebook shares of 115 known pro-Trump fake stories and 7.6 million of 41 known pro-Clinton fake stories. In addition, false stories often emerge surrounding natural disasters, such as the Japan earthquake in 2011 [110] and Hurricane Sandy in 2012 [37], that are intended to cause increased panic and disorder, or surrounding specific individuals and public figures, such as the death hoax of Singapore's first prime minister Lee Kuan Yew [20]. Another severe instance of the impact of fake news was the infamous "Pizzagate" incident wherein physical violence ensued as a result of fake stories circulated online [29, 66]. It is clear that the magnitude, diversity, and substantial dangers of false information in circulation are a genuine cause for concern.

The topic of fake news has not only received tremendous public attention but has also drawn increasing attention from the academic community. In this regard, there have been attempts to survey and summarize the literature on fake news detection. In one of the first surveys in the area, Shu et al. [101] illustrates an intriguing connection between fake news and social and psychological theories, which suggest that humans tend to seek, consume, and believe information that is aligned with their ideological beliefs, which often results in the perception and sharing of false information as true in communities of like-minded people. We note that the past year has witnessed a lot of new work related to fake news, which extends beyond the works surveyed in [101]. Another survey by Zubiaga et al. [132] study a related problem of rumor detection, which differs subtly from fake news detection in that it seeks to distinguish between verified and unverified information, wherein the unverified information may turn out to be true or false or may remain unresolved. The primary objective of this survey was to provide an architecture for developing a rumor classification system. In turn, the work summarizes some of the peripheral issues with identifying and tracking rumors, such as data collection and tools exposed by various social media platforms like Twitter and Facebook for the same, as well as an extensive list of applications and ongoing research projects toward this end. The literature surveyed in [132] is specifically restricted to that which is known to have been applied in the context of rumors. However, rumor classification and fake news detection are closely related in characteristics and techniques. In this work, we attempt to bridge the gap by comprehensively summarizing related techniques from closely related contexts. Another survey by Kumar and Shah [54] addresses a broader scope of false information on the web. In particular, aside

[1] http://www.theguardian.com/world/2017/jan/09.

from fake news on social media, the survey discusses fake reviews on e-commerce platforms and hoaxes in collaborative platforms such as Wikipedia. The nature and characteristics, and therefore techniques for detection of fake reviews and Wikipedia hoaxes, differ significantly from fake news on social media; for instance, fake reviews are opinion based where there is no single truth value, information does not propagate over a network and has specific characteristics such as ratings that are not applicable to fake news. Although these are extremely important topics in their own right, much like the detection of scam email [98], fake followers [23], or false web links [57], we chose to keep the discussion focused on fake news and provide detailed discussions of the techniques and their limitations. Moreover, Kumar and Shah [54] do not discuss some important existing works in fake news detection [61, 63, 89, 96, 118] and mitigation, such as [27, 48, 112], that are important to understand the advances in this domain. In particular, the purpose of our survey is fourfold:

(1) We address both detection and mitigation methods, including intervention-based techniques, to allow us to understand the problem in an end-to-end manner and go beyond treating the problem as a classification task.
(2) We focus on the challenges of developing computational methods to tackle fake news and present and compare existing methods that have demonstrated significant progress in dealing with those challenges.
(3) Existing surveys lack a comprehensive coverage of available fake news detection datasets. We consolidate a large comprehensive list of datasets and summarize their characteristic features—we hope this will aid researchers in selecting the right dataset, facilitate the collection of new datasets, and advance development of methods that can unify different information sources.
(4) Additionally, the rapid advancement of research around fake news necessitates the consolidation of recent works and advances and outlining concrete directions for future research.

Fake news identification and mitigation is a critical and socially relevant problem that is also technically challenging for a variety of reasons. In the following sections, we first formally define and characterize fake news and describe the problem and associated challenges. We then present a comprehensive overview of the existing techniques applicable to detection and mitigation, along with a discussion of the key limitations and recent advances in various methods. Further, to facilitate researchers with rigorous evaluation and comparison, we compile a list of existing/available datasets around fake news detection and summarize their characteristic features. To conclude, we enumerate a list of challenges and open problems that outline promising directions for future research.

2 DEFINITION, NATURE, ASSOCIATED CHALLENGES, KEY PLAYERS

In this section, we start by defining fake news and describing the nature and characteristics of the problem, as well as the key players involved, and the roles they play in information dissemination, moderation, and consumption. Further, we discuss the main challenges associated with fake news identification and mitigation and outline the necessary goals of any system that aims to address these challenges in an end-to-end, complete, and practically effective manner.

2.1 Definition

The usage and meaning of the term *fake news* has evolved over time. A Google Trends Analysis of the term reveals a sudden burst in popularity around the time of the 2016 US presidential election.[2]

[2] https://trends.google.com/trends/explore?date=2013-12-06%202018-01-06&geo=US&q=fake%20news.

Although originally used to reference false and often sensational information disseminated under the guise of news reporting,[3] the term has evolved and become synonymous with the spread of *false information* [22]. Fake news has generally been defined as "a news article that is intentionally and verifiably false" [2, 101] or "information presented as a news story that is factually incorrect and designed to deceive the consumer into believing it is true" [35]. However, the existing definitions are narrow, restricted either by the type of information or the intent of deception, and do not capture the broader scope of the term based on its current usage. Therefore, we define fake news as follows:

Definition 2.1. A news article or message published and propagated through media, carrying *false* information regardless the means and motives behind it.

This definition allows us to capture the different types of fake news identified in [119] that can be differentiated by the means employed to falsify information, such as fabricated content (completely false), misleading content (misleading use of information to frame an issue), imposter content (genuine sources impersonated with false sources), manipulated content (genuine information or imagery manipulated to deceive), false connection (headlines, visuals, or captions that do not support the content), and false context (genuine content shared with false contextual information). The definition also allows us to include different types of fake news identified by their motive or intent, such as malicious intent (to hurt or disrepute), profit (for financial gain by increasing views), influence (to manipulate public opinion), to sow discord (to create disorder and confusion), passion (to promote ideological biases), and amusement (individual entertainment) [126]. We can also subdivide false information by intent as misinformation and disinformation. Misinformation refers to unintentionally spread false information that can be a result of misrepresentation or misunderstanding stemming from cognitive biases or lack of understanding or attention, and disinformation refers to false information created and spread specifically with the intention to deceive [54]. Another type of information that might be closely connected to fake news is satire—satire presents stories as news that might be factually incorrect, but the intent is not to deceive but rather to call out, ridicule, or expose behavior that is shameful, corrupt, or otherwise "bad" [35]. The intent behind satire seems legitimate enough to exclude it from the definition, however, [119] does include satire as a type of fake news when there is no intention to cause harm but it has potential to mislead or fool people. Also, [35] mentions that there is a spectrum from fake to satirical news that they found to be exploited by many fake news sites, which used disclaimers at the bottom of their webpages to suggest they were "satirical," even when there was nothing satirical about their articles, to protect them from accusations about being fake. Thereby, our definition must include articles that are falsely posed as satire, as well as satirical articles that can potentially mislead, and exclude others that do not fall in this area. Additionally, we disambiguate the terms *hoax* and *rumor*, which are closely related to fake news. A hoax is considered to be a false story used to masquerade the truth, and, by the traditional definition, fake news can be seen as a form of hoax usually spread through news outlets [132]. The term *rumor* refers to unsubstantiated claims that are disseminated with the lack of evidence to support them. This makes them very similar to fake news, with the main difference being that they are not necessarily false and may turn out to be true [132]. Rumors originate from unverified sources but may later be verified as true or false or remain unresolved. Thereby, our definition can be seen as naturally encompassing hoaxes and false rumors.

[3] As defined in the Collins English Dictionary.

2.2 Nature/Characteristics

The definition of fake news is not the only thing that has changed with time. With the growth of computer-mediated communication through social media, we can see that the nature and characteristics of the problem have also evolved. Hence, we start by reviewing the literature from sociology and psychologically that explain the existence and spread of fake news at both an individual and social level.

2.2.1 Individual Level. The inability of an individual to accurately discern fake from true news leads to the continued sharing and believing of false information in social media. YouGov [124] found in a survey of 1,684 British adults who were shown six individual news stories, three of which were true and three of which were fake, that only 4% were able to identify them all correctly. The inability to discern has been attributed to cognitive abilities and ideological biases. Pennycook and Rand [81] identified a positive correlation between propensity for analytical thinking and the ability to discern false from true information. In addition, Allcott and Gentzkow [2] observed that people who spend more time consuming media, people with higher education, and older people had more accurate perceptions of information. The results were statistically significant in a survey of 1,208 US adults. Another study examined the impact of cognitive ability on the durability of opinions to find that individuals with lower cognitive ability adjusted their assessments after being told that the information given was incorrect but not nearly to the same extent as those with higher cognitive ability [93]. Besides cognitive abilities, ideological priors play an important role in information consumption. Naive realism (individuals tend to more easily believe information that is aligned with their views), confirmation bias (individuals seek out and prefer to receive information that confirms their existing views), and normative influence theory (individuals choose to share and consume socially safe options as a preference for social acceptance and affirmation) are generally regarded as important factors in the perception and sharing of fake news [101]. The survey by Allcott and Gentzkow [2] also found with statistical significance that people (Democrats and Republicans) are, respectively, 17.2% and 14.7% more likely to believe ideologically aligned articles than they are to believe nonaligned articles, although the differences in magnitude across the two groups (Democrats and Republicans) are not statistically significant. These individual vulnerabilities have been exploited to successfully disseminate fake information. Higgins [41] declared this era as an era of "post-truth," wherein objective facts are less influential in shaping public opinion than appeals to emotions and personal beliefs.[4]

2.2.2 Social Level. The nature of social media and collaborative information sharing on online platforms provides an additional dimension to fake news, popularly called the *echo chamber* effect [101]. The principles of naive realism, confirmation bias, and normative influence theory discussed above in Section 2.2.1 essentially imply the need for individuals to seek, consume, and share information that is aligned with their own views and ideologies. As a consequence, individuals tend to form connections with ideologically similar individuals (social homophily), and algorithms tend to personalize recommendations (algorithmic personalization) by recommending content that suits an individual's preferences, as well as by recommending connections to similar individuals to befriend or follow. Both social homophily and algorithmic personalization lead to the formation of echo chambers and filter bubbles, wherein individuals get less exposure to conflicting viewpoints and become isolated in their own information bubble [33]. The existence of echo chambers can improve the chances of survival and spread of fake news that can be explained by the phenomena of social credibility and frequency heuristic, where social credibility suggests that people's perception of credibility of a piece of information increases if others also perceive

[4]As defined by the Oxford English Dictionary, 2016.

Fig. 1. Nodes represent a sample of about 900 Twitter accounts that participated in the conversation about the US presidential election between June and mid-September 2016, and edges represent mutual-follower relationships between these accounts. Nodes are sized according to relative PageRank importance in the depicted network and colored according to inferred political ideology (red: right-leaning, blue: left-leaning, white: unsure) [34]. Reprinted with permission from [34].

it as credible, and frequency heuristic refers to the increase in people's perception of credibility with multiple exposures to the same information [101]. Gillani et al. [34] examined a network of 1.1M Twitter users who participated in conversations about the US 2016 presidential election between June and September 2016 and observed the existence of echo chambers and polarization of views based on the political orientation (democratic and republic) as visualized in Figure 1 that shows the follower relationships (edges) between users marked by the color of their political orientation (which is inferred from a user's profile and tweet contents as per [114]). Quattrociocchi et al. [90] similarly studied polarization by scientific vs. conspiracy narratives in Facebook users and observed that although both types of information were consumed similarly overall, 76.79% of all users who interacted on scientific pages and 91.53% of all users who interacted with conspiracy posts had 95% of their likes on either science or conspiracy posts. In addition, polarized users had a higher number of friends who displayed the same behavior, and higher edge homogeneity, measuring the similarity between friends in the network, suggested that it is highly unlikely that information would propagate across the different groups.

Fake news diffusion patterns on social media have often been studied to identify the characteristics of fake news that can help differentiate it from true news. Fake news detection works essentially rely on exploiting these differences to classify information based on its veracity. Most existing works primarily target the classification as a binary classification task (fake/true, rumor/non-rumor, hoax/non-hoax) or as a multi-class classification task (true/mostly true/half true/mostly false/false, unverified rumor/true rumor/false rumor/non-rumor). The main difference in different task settings is due to different annotation schemes or applications contexts in different datasets. Usually the datasets are collected from annotated claims on fact-checking websites such as `PolitiFact, Snopes`, and others and therefore reflect the labeling scheme used by the particular fact-checking website/organization. In some instances, credibility scores are provided based on human annotator judgments instead of class labels. A detailed description of different datasets and their annotations is provided in Table 2 when we summarize and discuss existing datasets in Section 7. In the remainder of this section, we examine different characteristics of fake news that are utilized for detection. We can identify three primary characteristics relevant for fake news detection, namely the source/promoters of the information, the information content, and the user responses it receives on social media.

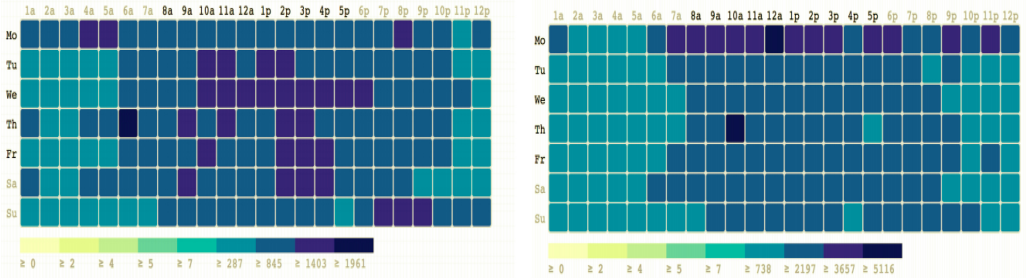


Fig. 2. The heatmap of the day of week vs. hour of tweets posted related to fake news (left) and real news (right). Darker colors indicate greater posting activity in a particular time slot [100]. Reprinted with permission from [100].

(1) **Source/promoters.** Zimdars [131] maintains a list of web addresses of fake news websites, several of which are modified names of true news websites, such as "abcnews.com.co" and "washingtonsblog.com." The use of such misleading domain names are a particular characteristic of fake news sources that individuals must learn to recognize and be attentive about. However, there are two caveats to filtering information based solely on these lists. One is that not all articles from these sources are fake and the other is that the list can never be exhaustive. In particular, the use of online resources and social media make it very easy, convenient, and inexpensive to create bots and register new accounts or domains [54]. Bots are fake or compromised accounts controlled by humans or programs to introduce and promote information on social media. Subrahmanian et al. [108] found a significant overlap between follower and followees of bot accounts that spread false information, which helps to engineer virality and credibility of posts and inflate the social status of the bot accounts. Davis et al. [24] found that on Twitter large number of bot accounts were responsible for accelerating the speed of both true and false information roughly equally. Shu et al. [100] randomly sampled 10,000 users who posted fake and real news on Twitter and used the bot detection algorithm by [24] to find that almost 22% of users involved in fake news are bots, while only around 9% of users are predicted as bot users for real news. They also investigated the temporal patterns of account activities by capturing the relationship between number of posts on Twitter at different times and days of the week for posts related to fake and true news and observed that periods of high posting activity included odd hours when people are generally inactive, suggesting the existence of social bots [100]. The temporal patterns observed by Shu et al. [100] are shown in Figure 2. Last, analyzing how source and promoters of fake news operate over the web across multiple online platforms, Zannettou et al. [126] found that false information is more likely to spread across platforms (18% appearing on multiple platforms) compared to true information (11%), with Reddit to Twitter to 4chan being the most common direction of information flow.

(2) **Information content.** The content of the information being spread is primarily what needs to be classified as true or fake. Horne and Adali [43] identified certain characteristics that differentiate fake news contents from true news contents by studying the textual characteristics of articles from various fake news websites and contrasting them with articles from reputed journalistic websites. Their findings suggest that titles of fake news articles are longer, have more capitalized words, and use fewer stop words, and the body content of fake news articles are shorter and repetitive and have fewer nouns and

analytical and technical words. Pérez-Rosas et al. [82] found that fake news articles contained more social words, with more verbs and temporal words, suggesting that the text tends to be focused on the present and future, instead of being more objective and factual. Other textual cues include self-reference, negation statements, complaints, and generalizing items [116]. An analysis by Newman et al. [74] revealed that deceptive stories had lower cognitive complexity, fewer exclusive words, more negative emotion words, and more motion (action) words. Silverman [102] observed that 13% of 1,600 news articles had incoherent headlines and content and used declarative headlines paired with article bodies that are skeptical about the claim in the headline [54].

(3) **User responses.** User responses on social media provide auxiliary information that is extremely beneficial for fake news detection. They are known to provide stronger signals for detection than the information content, mainly because user responses and propagation patterns are harder to manipulate than the information content and oftentimes contain obvious information about the veracity [100]. This secondary information in the form of user engagements (likes, shares, replies, or comments) contains rich information captured in the *propagation structure* (tree), which indicates the path of information flow, temporal information in *timestamps* of engagements, textual information in *user comments*, and *user profile* information by the user involved in the engagement. Zubiaga et al. [132] note that it is possible to differentiate user responses by their stance, wherein the most commonly used categorization uses the following four types, namely supporting, denying, querying, and commenting (which can be neutral or unrelated). They also observe that the nature of user responses varies depending on the stages of propagation, where, in the case of rumors, it was observed that majority of the users support true rumors and a higher number deny false rumors when the entire lifecycle of the rumor was considered, whereas by studying only the early reactions to rumors, it was found that users showed a tendency to support rumors independent of their veracity, which is suggestive of the fact that users have problems determining veracity in the early stages. Qian et al. [89] similarly found that fake news tends to receive more negative and questioning responses than true news. Another analysis of true and fake news captured the variation of sentiments in user replies, wherein it was observed that sentiment in replies to true news tended more toward neutral, in contrast to that for fake news, which tended more toward negative sentiments [100]. Friggeri et al. [31] also found from an analysis of user responses on Facebook that user responses change once false information is debunked, with a 4.4 times increase in deletion probability, even in early stages of the propagation.

2.3 Information Exchange Process

There are several different entities (individuals and organizations) that are *simultaneously* at play when it comes to the dissemination, moderation, and consumption of fake news through social media, which makes the problem of identification and mitigation more complex and involved. We discuss each part of the information exchange process.

2.3.1 Dissemination. A noticeable shift in information dissemination channels from traditional forms of journalism to online social media has been observed [78, 83]. In a survey of 3,000 journalists, 20%, responded that they thought that social media spelled the death of journalism [78]. Social media sites have become the popular form of dissemination due to growing ease of access and popularity of computer-mediated communication. Fake news can have a larger impact through social media due to the large scale and reach of social media and the ability to collaboratively share content [40].

2.3.2 Moderation. While in traditional forms of journalism the responsibility of content creation rests with the journalist and the reporting organization, moderation in social media varies substantially. In 2017, Germany passed the NetzDG (Network Enforcement) Act[5] to enforce removal of fake news within 24 hours (or up to a week depending on the complexity) by social media platforms with more than two million users. In UK, the parliament launched an investigation into how fake news is threatening modern democracy [38]. Nevertheless, the question of distribution of responsibility remains unresolved. A recent survey by Barthel et al. [5] determined that Americans collectively assign a fairly high and equal amount (45%) of responsibility to the government/politicians; social media platforms/search engines; and, last, members of the public; and, specifically, 15% hold all three responsible, while 27% hold two and 31% hold one of three responsible [5].

2.3.3 Consumption. Information is primarily consumed by the general public or society, which is a growing body of social media users. In a 2018 survey on social media usage with 2002 US adults surveyed, [104] established that 68% Americans use of Facebook, 73% use Youtube, and between 20 and 40% use other social media platforms such as Twitter and Instagram, and the numbers have grown almost 10-fold since 2005 [83]. Smith and Anderson [104] also established that 74% of the Facebook users use the site daily, with 51% using it several times a day; and a large fraction of frequent users lie in the 18–29 range. This growth in information consumption through social media adds to the risks of fake news causing widespread damage.

2.4 Key Players

We now consider a more subtle aspect that usually characterizes fake news, that is, the intent to deceive. In this light, we discuss the different roles that entities (individuals and organizations) play when it comes to dealing with fake news.

2.4.1 Adversary. Malicious individuals and organizations with a political or social agenda often pose as ordinary social media users using social bots [6] or actual accounts and can act as the source as well as promoters of fake news. Such accounts are known to also indulge in group behavior wherein groups of such accounts coordinate and share the same set of fake news articles [96]. The general characteristics of sources and promoters of fake news have been discussed in Section 2.2.

2.4.2 Fact-checker. In an attempt to combat the growing amount of false information, various *fact-checking* organizations, such as `Snopes` and `Politifact`, have been initiated to expose or confirm news stories. While these organizations are based on "fact-checking journalism" that relies on human verification, more desirable automated technological solutions have been proposed by technological companies like `Factmata`, which aim to provide fake news detection solutions to businesses and consumers and assign credibility scores to web content using artificial intelligence. Various other automated solutions in the form of plug-ins and applications such as `BS-Detector` and `CrossCheck` provide similar automated fact-checking services. An exhaustive list of fact-checking applications is provided in [132].

2.4.3 Susceptible. Fake news affects a wide range of individuals and organizations based on the motive. We summarized different motives or intents behind the spread of fake news in Section 2.1. For instance, reputable institutions and individuals might be susceptible to the attacks of fake news that is intended to sway public opinion about them, such as what was witnessed during the US 2016 presidential election [2]. Other consequences can even prove to be an increased risk to the entire world. Roozenbeek and van der Linden [94] noted that false information discrediting the

[5]https://www.bbc.com/news/technology-42510868.

seriousness of global warming can affect people's perception of climate change, posing a great risk to society and the world at large.

2.5 Challenges

The nature of the problem presents several challenges, which we summarize as follows.

2.5.1 High Stakes and Multiple Players. The World Economic Forum (2013) has ranked the spread of false information as one of the "top risks" the world is facing today. Moreover, the involvement of multiple entities and technological platforms increases the difficulty of studying and designing computational, technological, and business strategies, without compromising rapid and collaborative access to high-quality information.

2.5.2 Adversarial Intent. Malicious intent in content design and promotion increases the complexity of the problem. The content is designed to not only make it harder for humans to identify fake news by exploiting their cognitive abilities, emotions, and ideological biases as discussed in Section 2.2.1 but also to make it more challenging for computational methods to detect fake news. Shu et al. [100] evaluated the performance of several different methods on two datasets from PolitiFact and GossipCop and reported a maximum detection accuracy of 69% and 79.6%, respectively, even when using both article contents and social context, i.e., user responses to the article on Twitter.

2.5.3 Public Susceptibility and Lack of Awareness. To raise public awareness, numerous articles and blogs have been written that provide tips on differentiating truth from falsehood. For example, award-wining journalist Laura McClure highlighted five important questions to ask yourself when trying to determine whether a news article is true or fake. In one of the questions, McClure asked the reader to consider how an article makes them feel, citing that fake news is often "designed to make you feel strong emotions."[6] Articles such as McClure's are informative and effective for individuals; however, they do not provide a scalable and systematic solution to the problem.

2.5.4 Propagation Dynamics. The dynamic nature of the process of fake news propagation through social media further complicates matters. False information can easily reach and impact a large number of users in short time [31, 89]. Friggeri et al. [31] studied rumor cascades on Facebook and found that information is readily and rapidly transmitted, even when it is of dubious veracity. Fact-checking organizations like `Snopes` and `Politifact` cannot keep up with the propagation dynamics as they require human verification, which inhibits a timely and cost-effective response [48, 96].

2.5.5 Constant Change. Fast-paced developments in the world pose additional challenges to knowledge-based systems that need to dynamically retrieve and update their state based on newly emerging facts [87]. Fact checking is ultimately essential for reliably identifying fake news. For example, while scandals are not uncommon among celebrities, when a new scandal about a celebrity comes out, without enough extra knowledge, it is very difficult to tell whether it is fake or not. Potthast et al. [87] note that style-based fake news detection, i.e., differentiating fake and true news by an analysis of writing styles, is simply an alternative used due to the unresolved challenges in automating fact checking from knowledge bases.

[6] http://blog.ed.ted.com/2017/01/12/how-to-tell-fake-news-from-real-news/.

2.6 Requirements/Goals

Existing works have demonstrated significant progresses toward alleviating some of the challenges in fake news detection and mitigation. However, there is still room for improvement before the problem can be addressed in more effective ways. We suggest few requirements that are of interest in developing solutions to tackle fake news.

2.6.1 Balancing Aggressive and Non-aggressive Moderation. Identification and mitigation techniques that require very aggressive moderation by social media platforms can hurt the social media platform [27]. Thereby, it seems necessary to design strategies that still effectively mitigate the problem of fake news but without restricting rapid access to high-quality information and collaborative information sharing. We believe that it is also important that the moderation strategies do not further introduce more confusion and distrust into the environment. Fake news has already resulted in people becoming skeptical of even true information, which hurts the value of social media as a platform for information sharing. A survey of 1,002 US adults [5] found that (64%) people say that fake news has caused a great deal of confusion about the basic facts of current issues and events, 24% say it has caused some confusion, and 11% say not much/no confusion.

2.6.2 End-to-end Solutions. Reliable and timely detection of fake news should be accompanied by computational methods to intervene and prevent further spread of news that is confirmed as fake. In addition, it would be desirable to design interventions that can provide quantifiable measures of the impact of the intervention effort in terms of the exposures to fake and true news [27].

2.6.3 Balancing Timeliness vs. Detection Accuracy. Early detection and mitigation are critical goals of any effective system. However, the available information for detection increases as time progresses, with only the content of the article being available at the start, followed by increasing user responses as propagation continues [89]. Most existing methods rely on content only or on user responses only or do not utilize responses incrementally. Detection systems must aim to utilize incrementally available information to trade off confidence in detection accuracy vs. timeliness of the detection and mitigation effort.

2.6.4 Prioritization and Cost-effectiveness. The ability to optimally decide which contents to fact-check at what time can equip the system to provide better responses by being able to quickly remove false information that can have a potentially larger and faster impact than those that might have a negligible or slower impact if allowed to propagate further in time [48, 79]. Also, human involvement in fact checking increases not only the delay but also the cost of intervention, which necessitates the need for prioritization of information to manually fact-check, until reliable automated methods can be sought [79].

2.6.5 Robustness, Scalability, and Interpretability. The high stakes and consequences of fake news necessitates the need for reliability in detection. Mistakenly removing true information from the platform, or not detecting and removing potentially viral false information, would become problematic in practice. To move from manual and semi-automated solutions to fully automated ones will not be possible without robust and also interpretable predictions. Popat et al. [84] emphasized transparency and interpretability in credibility assessment systems and built CredEye, which verifies an input claim and provides the probability of truth and falsehood along with extracted evidence (from fact-checking or trusted news websites) with words that support the claim highlighted in green, refuting the claim highlighted in red, and words overlapping with the claim provided in yellow and the intensity of colors reflects the word's importance for the assessment (based on feature weights from the learned classifier).

Table 1. Categorization of Existing Methods

Content-Based Identification	Feedback-Based Identification	Intervention-Based Solutions
Cue and feature methods	Hand-crafted features	Mitigation strategies
Linguistic analysis methods	Propagation pattern analysis	Identification strategies
Deep learning content-based	Temporal pattern analysis	
	Response text analysis	
	Response user analysis	

2.6.6 Evolution and Up-to-date Fact Checking. The adversarial nature of fake news necessitates that the system should be able to dynamically adapt to the changing strategies of adversarial opponents. Specifically, adversaries create new accounts and throw-away accounts to promote fake news and avoid detection [54], as can be seen in an instance of the US election when a potentially large number of fake accounts were created to influence the election and were found to initiate coordinated attacks with specific political agendas.[7] In addition, increased sophistication in adversarial techniques are speculated in terms of both fake content creation [54], as well as in strategies to promote fake content. Ruchansky et al. [96] observed from behaviors of suspicious users on Twitter that these users had more similar engagement patterns toward true and fake news than on Weibo, which could demonstrate an increased sophistication in fake content promotion on Twitter.

3 OVERVIEW OF METHODS

We divide the existing work into three types. The first type is fake news identification using content-based methods that classify news based on the content of the information to be verified. The second type is identification using feedback-based methods that classify news based on the user responses it receives on social media. Last, the third type is intervention-based solutions that provide computational solutions for *actively* identifying and containing the spread of false information and methods to mitigate the impact from exposures to false information. Each category is further divided based on the type of existing methods, as shown in Table 1.

4 CONTENT-BASED IDENTIFICATION

In this section, we give an overview of content-based approaches to fake news detection. The underlying basis of content-based detection is that textual content in fake news differs from that in true news in some quantifiable way. The use of language cues to determine veracity was first motivated by work in applied psychology for evaluation of eyewitness testimonies [113]. Language cues can be exploited with traditional hand-engineered feature-based methods, linguistics-based methods, and more advanced deep learning methods that bypass feature engineering for content analysis.

4.1 Cue and Feature-based Methods

Cue and feature-based methods can be employed to distinguish fake news contents from true news contents by designing a set of linguistic cues that are informative of the content veracity. Several different cue sets have been proposed in literature.

4.1.1 Scientific Content Analysis (SCAN). One of the earliest works on examining the use of linguistic cues for fake news detection was by Driscoll [26], which studied transcripts or written

[7] https://www.nytimes.com/2017/09/07/us/politics/russia-facebook-twitter-election.html.

statements made by individuals in a criminal investigation. To determine the credibility of the information given by suspects in such statements, they utilized an approach called Scientific Content Analysis (SCAN), which was proposed by the polygraph examiner Sapir [99], based on his experience with subjects of polygraph examinations. SCAN consists of cues related to deception detection. The cues include content and structural criteria such as lack of connection between paragraphs, lack of conviction or memory, denial of allegations, missing and out-of-order information, use of emotive words, objective or subjective words, pronouns, first person, singular, and past-tense verbs. While the examination by Driscoll [26] revealed positive results in differentiating true from false statements using SCAN, later works have found them to be ineffective based on more rigorous evaluations, finding no significant differences between true and fabricated statements concerning these criteria [9, 72].

Limitation. Although intuitive, SCAN is set of subjective criteria lacking enough supporting evidence; moreover, the approach requires the use of trained professionals to analyze the statements for veracity, making it difficult to automate.

4.1.2 Linguistic-based Cue Set (LBC). In an attempt to reduce human involvement in deception detection, one of the pioneering automated text analysis methods was by Fuller et al. [32]. Fuller et al. [32] consolidated several linguistic-based cues and previously proposed cue sets. The first cue set included was the Zhou/Burgoon set [130] comprising 14 linguistic-based cues that were found effective for deception detection and included the percentage of first-person singular and plural pronouns in the text, average word length, verb quantity, sensory ratio, spatial and temporal ratio, and imagery. The second set of cues was derived from deception constructs drawn from deception theories [12, 13] that included sentence and word quantity, activation, certainty terms, generalizing terms, imagery, and verb quantity. The third cue set was a comprehensive set of 31 cues created by including the first two cue sets along with additional Linguistic Inquiry and Word Count–(LIWC) [80] based cues utilized by previous studies [10, 74] and included lexical diversity, modal verbs, passive verbs, emotiveness, exclusive terms, and redundancy. The final cue set was a refined cue set determined by feature selection to identify the most important of the 31 cues in the comprehensive set. To determine the relative importance of cues, Fuller et al. [32] utilized three different classifiers, that is, neural network, decision tree, and logistic regression, on statements of witnesses in official investigations, wherein the input feature vector to the classifier consisted of the normalized frequency of cues appearing in the text. The eight cues obtained through feature selection were third-person pronouns, content word diversity, exclusive terms, lexical diversity, modifiers, sentence quantity, verb quantity, and word quantity.

Limitation. The drawback of using linguistic-based cue sets is the lack of generalizability across topics, languages, and domains. Ali and Levine [1] showed that a linguistic cue set designed for one situation may not be suitable for another situation due to language variations, e.g., a cue set designed for accounting [58] or police interrogation [86] may differ significantly.

4.1.3 Other Variants. Later works have explored refined hand-crafted cue sets more closely targeted toward the problem of fake news detection. Rubin et al. [95] analyzed several text analysis-based features, such as the number of punctuation marks and the sentiment of the text. Zhao et al. [128] proposed different regular expressions to capture enquiry and correction patterns in posts on social media, which are indicative of rumors such as "is (that/this/it) true," which capture enquiry and "(that/this/it) is not true," which capture correction. They also included some platform-specific features such as counts of "hashtags" and "mentions" in posts on Twitter, as well as the ratio of enquiry or correcting posts in a cluster of posts having high textual similarity, i.e., posts discussing similar content. Others works that have proposed cue sets designed specifically for certain types

of social media platforms include [15, 36, 69] for Twitter, [55] for Wikipedia, and [17] for click-bait websites.

Limitation. Exhaustive enumeration of regular expression patterns is non-trivial and requires significant effort. Furthermore, identifying relevant platform-specific features for a large variety of social media platforms is also challenging and reduces applicability of the method across platforms and domains.

Limitations of cue and feature-based methods. To summarize the limitations, variations in linguistic cues implies that a new cue set must be designed for a new situation, making it hard to generalize cue and feature engineering methods across topics and domains. And such approaches thereby involve more human involvement in the process to design, evaluate, and utilize these cues for detection.

4.2 Linguistic Analysis-based Methods

While manual cue selection is intuitive and interpretable, it is often specific to the setting being considered and does not generalize easily to other settings. In an attempt to make cue-based models more general, methods based on linguistic analysis were proposed. Linguistic analysis-based methods, like cue-based methods, can be applied to distinguish fake from true news by exploiting differences in writing style, language, and sentiment. Such methods do not require task-specific, hand-engineered cue sets and rely on automatically extracting linguistic features from the text. We discuss three linguistic analysis-based methods that are applied to fake news detection.

4.2.1 N-gram Approach. The most effective linguistic analysis method applied to fake news detection is the *n-gram* approach [67, 76, 77]. n-grams are sequences of n contiguous words in a text, constituting words (unigrams) and phrases (bigrams, trigrams) and are widely used in language modeling and text analysis.

Approach. Mihalcea and Strapparava [67] proposed the use of n-grams for lie detection. They constructed datasets using crowd-sourcing that constituted statements of people lying about their beliefs on topics such as abortion and death penalty/lying about their feelings on friendship. Mihalcea and Strapparava [67] wanted to determine how the texts differed and whether n-grams analysis was enough to differentiate lies from the truth. They trained Naive Bayes and Support Vector Machine (SVM) classifiers with inputs being the term frequency vectors of n-grams in the texts after tokenization and stemming but without stop word removal. Interestingly, the classification accuracy was about 70% in identifying people's lies about their beliefs and 75% in identifying lies about their feelings. A fine-grained analysis of word usage revealed that in all the deceptive texts, connections to the self ("I, friends, self") were lacking and other human-related word classes ("you, other, humans") were dominant, indicative of the speaker's discomfort in identifying themselves with the lying statements. Also, it was found that words related to "certainty" were dominant in deceptive texts, which is probably explained by the need for the speaker to explicitly use truth-related words as a means to make their false statements more believable. [76, 77] provided similar n-gram classification analysis for deceptive reviews created by crowd-sourced workers on Amazon Mechanical Turk, who were asked to create fake positive reviews about hotels [77] and fake negative reviews about hotels [76]. Their analysis revealed that fake reviews contained fewer spatial words (location, floor, small), because the individual had not actually experienced the hotel and had fewer spatial detail available for the review. They also found that positive sentiment words were exaggerated in positive fake reviews compared to their true counterparts. A similar exaggeration was seen in negative sentiment words in fake negative reviews.

Limitation. Being a simplified approach, using n-grams alone cannot entirely capture finer-grained linguistic information present in the writing styles of fake news.

4.2.2 Part-of-Speech Tags. Apart from word-based features such as n-grams, syntactic features such as Part-of-Speech (POS) tags are also exploited to capture linguistic characteristics of texts. POS tags are obtained by tagging each word in a sentence according to its syntactic function, such as nouns, pronouns, adjectives; and several works have found the frequency distribution of POS tags to be closely linked to the genre of the text being considered, for example, medical consultations, committee meetings, and sermons each have their own distinctive pattern [7, 92]. Ott et al. [77] examined whether this variation in POS tag distribution also exists with respect to text veracity. They trained a SVM classifier using relative POS tag frequencies of texts as features on a dataset containing fake reviews. Ott et al. [77] obtained better classification performance with the n-grams approach but nevertheless found that the POS tag approach is a strong baseline outperforming the best human judge. A qualitative analysis showed that the weights learned by the classifier are largely in agreement with the findings of existing theories on deceptive writing such as in Reference [92], which suggests connections of deceptive opinions to imaginative writing comprising more verbs, adverbs, pronouns, and pre-determiners and truthful opinions to informative writing comprising more nouns, adjectives, prepositions, determiners, and coordinating conjunctions.

Limitation. The sole use of POS tags provides syntactic information alone and is weaker than word-based approaches that capture more information, inclusive of writing styles such as emotiveness inferred from words like excited, terrible, and so on.

4.2.3 Probabilistic Context Free Grammar. Later work has considered deeper syntactic features derived from Probabilistic Context Free Grammars (PCFG) trees [45]. A Context Free Grammar (CFG) tree represents the grammatical structure of a sentence with the terminal nodes representing words and intermediate nodes representing syntactic constituents such as verb, noun phrase, and so on. Depending on language construction ambiguities, a sentence can have multiple syntactic representations. PCFG enables disambiguation by associating a probability with each tree, where the probability of a tree is the product of the probabilities of all production rules in the tree. A production rule is represented as follows: $A \rightarrow \alpha$, where $A \in V$ and $\alpha \in (V \cup T)^*$, with V being the set of intermediate nodes and T being the set of terminal nodes. The production rule probability is statistically derived from the corpus.

Approach. Feng et al. [28] examined the use of PCFG to encode deeper syntactic features for deception detection. In particular, they proposed four variants when encoding production rules as features. The first variant includes only those production rules from the data that do not contain terminal nodes. The second variant includes all production rules derived from the dataset. The third and fourth variants modify the production rules to include grandparent nodes (i.e., parent of node A in rule $A \rightarrow \alpha$), with and without rules with terminal nodes, respectively. The feature vector is constructed using the tf-idf counts (normalized frequency) of production rules in the text. Feng et al. [28] trained an SVM classifier and found PCFG features used with n-gram features to be more beneficial than POS tags with n-grams in classifying fake texts from hotel reviews and lie detection datasets [67, 77]. Feng et al. [28] noted that of the four variants of production rules, the ones with the terminal nodes included were more powerful, since they contained word-based information in addition to syntax information. A qualitative analysis of the most discriminative syntactic constituents based on the classifier weights indicated the use of certain syntactic constituents such as indirect enquiry (WHADVP), verb phrases (VP), and subordinating conjunction clauses (SBAR) to be more frequent and important in fake texts as compared to true texts.

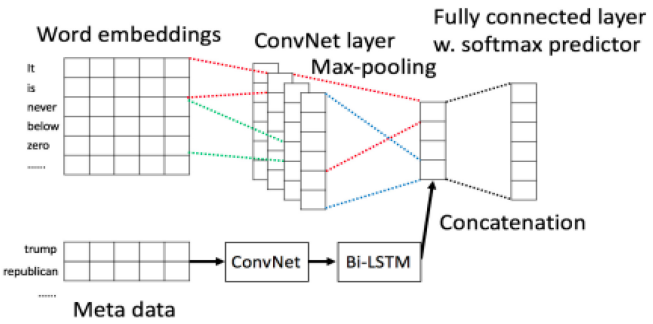


Fig. 3. CNN architecture used for fake news detection [118]. Reprinted with permission from [118].

Limitation. PCFG can be used to extract syntactic features from a sentence but is not powerful enough to capture context sensitive information from across sentences, thereby limiting its effectiveness in classification of longer fake news articles or texts.

Limitations of linguistic analysis-based methods. Even with word-based n-gram features combined with deeper syntactic features from PCFG trees, linguistic analysis methods, although better than cue-based methods, still do not fully extract and exploit the rich semantic and syntactic information in the content. n-gram approach is simple and cannot model more complex contextual dependencies in the text. Syntactic features used alone are less powerful than word-based n-grams, and a naive combination of the two cannot capture their complex interdependence.

4.3 Deep Learning Content–based Methods

Deep learning methods alleviate the shortcomings of linguistic analysis–based methods by automatic feature extraction, being able to extract both simple features and more complex features that are difficult to specify. Deep learning–based methods have demonstrated significant advances in text classification and analysis [8, 49, 123] and are powerful methods for feature extraction and classification with their ability to capture complex patterns relevant to the task.

4.3.1 Convolutional Neural Networks. Convolutional neural networks (CNN) [60] are generally used in natural language processing tasks such as semantic parsing [123] and text classification [49]. Wang [118] proposed the use of convolutional neural networks for content-based fake news detection. Wang [118] collected a dataset of short statements labeled based on the degree of falsehood by `PolitiFact`, a reputed fact-checking organization. Figure 3 shows the model architecture proposed in [118]. The model takes two inputs, the statement text and the speaker metadata information, such as political orientation, home state, and so on, available in the dataset. The text inputs are processed by a word-embedding layer to obtain continuous low-dimensional representations for each word in the text sequence. The output of this layer is processed by convolutional and max-pooling layers that generate the extracted feature representation. Similarly, the speaker metadata are also processed similarly by a different embedding and convolutional layer and a bidirectional LSTM [42] layer to generate its final extracted feature representation. The two representations are concatenated and fed to the classifier trained end-to-end with the other layers. Wang [118] utilized pre-trained word2vec embeddings [68] to warm start the text embeddings. These word embeddings are known to capture useful properties of word co-occurrences and contextual properties, and semantic relationships between words, trained on large corpora of unlabeled texts. Wang [118] observed better detection accuracy with the deep learning–based CNN model compared to using SVM and logistic regression and also found the inclusion of speaker metadata beneficial. Qian et al.

[89] also similarly demonstrated improved CNN performance over linguistic analysis-based methods such as the LIWC, POS, and n-gram approaches when classifying a collection of news articles as fake or true. In addition, to handle longer article texts, Qian et al. [89] suggested a variant of the CNN architecture called Two-Level Convolutional Neural Network (TCNN) that first takes an average of word-embedding vectors for words in a sentence to generate sentence representations and then represents articles as a sequence of sentence representations provided as input to the convolutional and pooling layers. Qian et al. [89] found the TCNN variant to be more effective than CNN in classifying the articles.

4.3.2 Other Variants. Recurrent neural network–(RNN) [97] based architectures are also proposed for fake news detection. RNNs process the word embeddings in the text sequentially, one word/token at a time, utilizing at each step the information from the current word to update its hidden state, which has aggregated information about the previous words. The final hidden state is generally taken as the feature representation extracted by the RNN for the given input sequence. A specific variant called Long Short-Term Memory (LSTM) [42], which alleviates some of the training difficulties in RNN, is often used due to the its ability to effectively capture long-range dependencies in the text and has been applied to fake news detection, similarly to the use of convolutional neural networks in several works [91, 117]. In another variant, LSTM has been applied to both the article headline and article text (body) in an attempt to classify the level of disagreement between the two for deception detection [19].

Limitations of deep learning–based methods. Even with sophisticated feature extraction of deep learning methods, fake news detection remains a challenge, primarily because the content is crafted to resemble the truth to deceive readers, and without fact checking or additional information, it is often hard to determine veracity by text analysis alone. A recent evaluation [100] benchmarking different methods on datasets of political statements and celebrity gossip news also confirms relatively low classification accuracy of 63% and 70% on the two datasets using purely content-based classification with CNN.

5 FEEDBACK-BASED IDENTIFICATION

In this section, we present and discuss feedback-based approaches for fake news detection. In content-based approaches, the text of an article is regarded as the primary source of information. However, rich secondary information in the form of user responses and comments on articles and patterns of news propagation through social media can likely be more informative than article contents that are crafted to avoid detection. These secondary information sources form the basis of the works discussed in this section.

5.1 Hand-engineered Features

Hand-engineered features proposed for fake news detection from feedback signals include various types of features, such as propagation pattern features, temporal pattern features, and text- and user-related features. Castillo et al. [15] designed a feature set to include user-based features (e.g., registration age and number of followers), text-based features (e.g., the proportion of tweets that have a mention "@"), and propagation-based features (e.g., the depth of the re-tweet tree) and used a decision tree model to classify news as fake using the designed feature vector. Variants of the above include other network-based features that are slightly extended or tailored to a particular context, such as the inclusion of temporal features [56] or geographic location and type of device (mobile or PC) from which the response was sent [122].

Limitation. Feature engineering allows us to incorporate diverse kinds of information, which is useful in this case. However, hand-engineered features are limited in terms of the generality and complexity of the feature space being captured.

5.2 Propagation Pattern Analysis

True and fake news spread through social media in the form of shares and re-shares of the source and shared posts, resulting in a diffusion cascade or tree, with the source post at the root. The path of re-shares and other propagation dynamics of fake and true news contents are utilized for fake news detection. In this section, we discuss works that specifically utilize propagation structures and patterns for fake news detection.

5.2.1 Propagation Tree Kernels. Propagation tree kernel methods exploit the differences in propagation patterns of two diffusion cascades (trees) as features for detection. Ma et al. [63] proposed to compare the similarity between propagation trees using *tree kernels* that were applied to syntactic modeling tasks in [21, 70, 127].

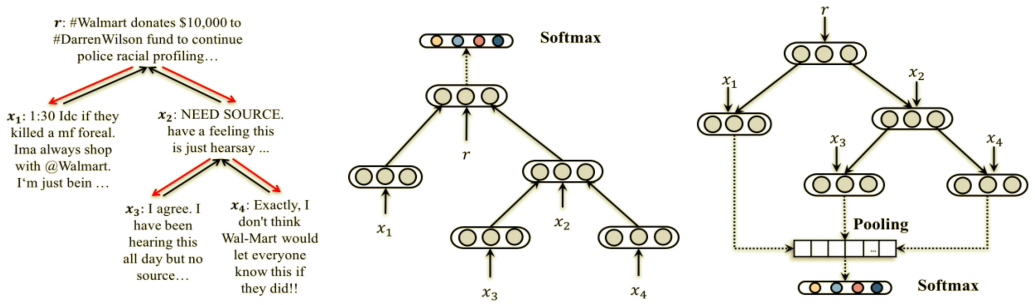
Approach. Ma et al. [63] defined a tree kernel that is utilized to compute the similarity between two trees as follows. Each node of the tree corresponds to a specific user engagement with the article and is associated with the timestamp of the engagement, textual information (user's comment/reply), and the user's metadata. These attributes form the feature set of the node. The similarity between two nodes (one from each propagation tree) is a function of the similarity between the node features and similarity over subtrees, computed recursively. Ma et al. [63] designed the tree kernel with different measures of similarity for different features. For textual features, they utilized Jaccard similarity as a measure of similarity over n-grams in the textual information $J(c_i, c_j)$. For the temporal similarity, they considered the absolute difference of the time lags $t = |t_i - t_j|$, where the time lag is basically the relative difference of timestamp of the engagement and the timestamp of the source post. For user metadata features, they considered the L2-norm (euclidean distance) between the user metadata feature vectors $\mathcal{E}(u_i, u_j)$. Combining the three metrics of similarity, the similarity between two nodes v_i and v_j is defined as follows:

$$f(v_i, v_j) = e^{-|t_i - t_j|}(\alpha\mathcal{E}(u_i, u_j) + (1 - \alpha)J(c_i, c_j)),$$

where v_i is a node from one propagation tree and v_j from the other; α is a constant; u_i, u_j are user metadata vectors; c_i, c_j are the textual information n-gram vectors; and t_i, t_j are the time lags of the engagements. Using the defined kernel, Ma et al. [63] computed the similarity between every pair of diffusion cascades in the dataset and used SVM with the defined tree kernel for classification, observing improved detection performance over methods that do not utilize propagation patterns. Wu et al. [120] similarly defined a random walk graph kernel [46] to calculate similarity between different propagation trees. The designed kernel function is composed of the random walk graph kernel over propagation trees and the standard radial basis function kernel over other features, paired with SVM, similarly to [63].

Limitation. Propagation tree kernel methods are computationally intensive (requiring the computation of pairwise similarities over all trees), which can prohibit their large-scale applicability to the task of fake news detection.

5.2.2 Propagation Tree Neural Networks. The limitations of propagation tree kernel methods have been addressed by more recent works [65] using deep learning methods. The advances proposed in Ma et al. [65] not only significantly reduce the computation costs and time but also manage to improve over the detection accuracy obtained by the tree kernel methods [63]. Ma et al. [65]



(a) Bottom-up/Top-down tree (b) Bottom-up RvNN model (c) Top-down RvNN model

Fig. 4. Tree structured recursive neural network architectures considered in [65]. Reprinted with permission from [65].

proposed the use of recursive neural networks generally used in syntactic and semantic parsing [105–107] to extract propagation features from diffusion cascades.

Approach. Ma et al. [65] design two recursive neural network models based on bottom-up and top-down tree-structured recursive neural networks. Similarly to the RNN mentioned in Section 4.3.2 that sequentially process each timestep (word) in the input sequence while aggregating information from previous timesteps (words), the recursive tree neural network sequentially processes each node in the tree by recursively traversing the tree structure in either the top-down or bottom-up manner based on the design. In the bottom-up approach, information is aggregated starting from leaf nodes and propagated upward, such that the final feature representation obtained at the root is representative of the complete propagation tree and is used for classification. The weights in the classification layer (output layer with softmax activation) and the parameters of the recursive network are trained end to end based on the classification loss. In the top-down approach, information flows over paths in the tree and the resulting representations at the leaves are combined to generate the extracted feature representation of the tree that is used by the classification layer. Figure 4 shows the proposed network architectures. Ma et al. [65] find that both architectures outperform propagation tree kernels and other baselines that do not utilize tree structured information in user engagements. They also find that the top-down approach performs better, which they believe is because it assumes the direction of information flow through the diffusion cascade.

Limitation. The sequential recursive nature of computations impacts training speed. In addition, since information will be aggregated and compressed over many nodes, in large cascades it might potentially not retain enough information from all nodes to obtain a good representation for the tree.

5.2.3 Propagation Process Modeling. Alternatively, extensions to models originating from epidemiology that model the spread of diseases [50, 73] have also been applied to characterize the propagation of fake news. Jin et al. [44] proposed a mathematical model called SEIZ to capture how people share news on social media. Jin et al. [44] consider a set of eight news stories spanning multiple topics such as politics, terrorist events, and so on, propagating on Twitter, of which four are true news and the rest are rumors, and examine whether the proposed model can characterize the studied cascades.

Approach. The proposed model SEIZ is defined by dividing users into four partitions: susceptible users ("S") that have not heard about the news yet, exposed users ("E") who have received the news and will share it in the future (with some delay), infected users ("I") that have shared it, and skeptic

users ("Z") that have heard about the news but do not share the news. A user can transition from one partition to another with a certain probability and at a certain rate, as defined by the model. The parameters of the model $\beta, \rho, b, \epsilon, p, l$ are the rates constants and probabilities of transitions between different partitions. p is the probability and β is the rate of transitions from S to I, where b is the rate of transitions from S to Z with probability l and S to E with probability $1 - l$. ρ is the rate of transition of users from E to I through contact with I, and ϵ is the rate from E to I by self adoption instead of contact. It is assumed that each partition has an associated size at time t, but the total size of all partitions combined remains constant over time (i.e., users only transit from one partition to another and the propagation dynamics are controlled by the parameters of the model). The size of a partition at a given time is obtained by solving a set of differential equations based on the initial size of partitions and the rate constants and probabilities, which form the parameters of the model. The parameters are estimated using nonlinear least squares fit on an observed cascade corresponding to true or fake news. Each step of the fitting process involves iteratively refining a set of parameter values by numerically solving the system of differential equations to minimize the difference between observed shares and the size of partition I based on the model $|I(t) -$ observed shares (t)|. Jin et al. [44] quantify a ratio (R_{SI}) as the ratio of influx into E from S to the outflow from E to I and suggest that it can potentially be used to detect rumors, where they found that higher R_{SI} is suggestive of true news based on the studied cascades. The ratio is defined as follows:

$$R_{SI} = \frac{(1 - p\beta) + (1 - l)b}{p + \epsilon},$$

where p, ρ, β, b, l, and ϵ refer to the model parameters mentioned earlier.

Limitation. The model parameters are not user specific and the same rate constants and probabilities are assumed for all users. Moreover, the model has not been specifically tested for fake news detection to validate its empirical success or computational feasibility.

5.3 Temporal Pattern Analysis

Differences in temporal dynamics of user engagements with articles can be leveraged for fake news detection. The length of time and the rate and intervals at which true and fake news spreads can differ. Extracting intricate interactions and variations along time explicitly is the focus of the work presented in this section.

5.3.1 Temporal Variation Features. Temporal modeling methods are applied to sequentially ordered responses $T_i = \{t_1, t_2 \cdots t_n\}$, in which each post t_i has a timestamp at which the interaction or user engagement with the article is recorded. In the simplest form, temporal features can be incorporated by taking both features in each time interval and the variation between features of two temporally adjacent intervals [62]. In particular, Ma et al. [62] consider three types of features, that is, text features (such as percentage of posts with exclamation or question marks, percentage of posts with hashtags or user mentions, average number of positive/negative emoticons), user features (such as percentage of verified users, average number of followers, registration age), and propagation features (such as average number of reshares, average number of comments). Ma et al. [62] divide every sequence of ordered posts into a fixed number of time intervals and each post in the sequence falls within some interval. Based on that, they compute the features for each time interval from aggregation of posts up to that interval. In addition, the variation of features between two adjoining time intervals is considered as the differences in the feature values divided by the interval length. The final feature vector is a concatenation of features from each interval together with the temporal variation features, which is paired with an SVM classifier in [62]. Ma et al. [62] observe reasonable performance improvements over methods that do not consider

temporal patterns and rely only on the overall statistics such as total number of reshares or time length of propagation and ignore their variations over time.

Limitation. One shortcoming of such a primitive feature variation model is that it requires handcrafted features that are further restricted to numerical features over which temporal variations can be considered.

5.3.2 Temporal Pattern with Recurrent Neural Networks. Intricate temporal variation differences between true and fake news can be automatically extracted with deep learning–based methods. RNN are effective in time-series modeling. The approach used in general for application to fake news detection is to divide the sequentially ordered engagements into discrete time intervals. Each interval is represented by a set of features. The ordered sequence of feature vectors is provided as input to the recurrent neural network. Each timestep is processed by the network by building on information processed in previous timesteps.

Approach. Ruchansky et al. [96] partition a given sequence of engagements using discrete time intervals at specific granularity. Timestamps of engagements are considered relative to the first engagement in the sequence. Each interval is represented by the aggregated features of engagements in that interval. Specifically, the feature vector has the following form, where η is the number of engagements in interval t, $\triangle t$ is the time between the current and previous non-empty interval, x_u is the average of user-features over users that engaged with the article during t, and x_τ is the textual content in engagements during t:

$$x_t = (\eta, \triangle t, x_u, x_\tau).$$

Together, η and $\triangle t$ provide a general measure of the frequency and distribution of the response an article received. The textual features x_τ can be generated in different ways. Ruchansky et al. [96], in particular, chose to learn x_τ from the raw text features using the doc2vec [59] model. Ruchansky et al. [96] trained an LSTM recurrent neural network [42] with the sequence of feature vectors x_t as input for classification of fake news based on the user engagements. They used two datasets, one with engagements on Twitter and the other on Weibo. Both datasets did not contain article contents and the classification is done solely using the user responses (engagements) available in the dataset, using the proposed LSTM model. The datasets also does not contain any propagation tree information, and, therefore, the engagements of a given article can only be ordered temporally. The model architecture proposed by Ruchansky et al. [96] also contains a second component. The first component we discussed here is to capture the temporal and textual patterns in user responses. The second component is to separately capture user characteristics and group behaviors and is explained in Section 5.5.2. Another earlier work by Ma et al. [61] also examined the use of recurrent neural networks to capture temporal patterns from engagements. The architecture by Ma et al. [61] does not include the second component capturing user characteristics. The feature representation chosen in the two works differs slightly in that Ma et al. [61] uses tf-idf vectors as text features, and temporal differences are not provided as features as in Ruchansky et al. [96] but are instead captured by sampling engagements at regular intervals from the time series. The slight differences in feature representations chosen by the two do not affect model performance. Ma et al. [61] also examined the performance of other RNN variants besides LSTM, such as Gated Recurrent Unit (GRU) [18] and a two-layered GRU network, and found the latter to have slightly better performance of all three variants.

Limitation. Cascades with too many user responses, would possibly require some sampling or pruning scheme to be efficiently represented, which may result in distortion of the temporal patterns.

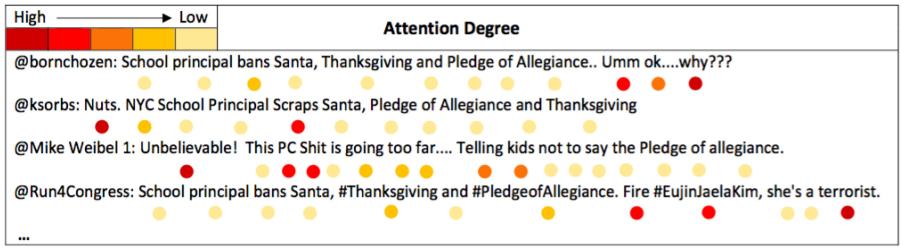


Fig. 5. Attention weights learned over different words in the sentences. Darker color is indicative of greater attention, making the corresponding words more relevant to the fake news detection task [16]. Reprinted with permission from [16].

5.4 Response Text Analysis

The text in user responses can be quite informative toward fake news detection. Even without access to the article content, response text can provide at least some insights about the article content. It is also likely that fake or dubious articles will receive more negative and questioning responses that can be leveraged for detection. In this section, we discuss methods that specifically exploit textual content in user responses.

5.4.1 Deep Attention. Work discussed in Section 5.3.2 utilized tf-idf features and doc2vec word embeddings to represent textual response features. Another variant of that work proposed by Chen et al. [16] focuses specifically on the extracted textual information by adding an attention mechanism into the LSTM architecture to capture representative words of fake news and are able to depict which portions of the text can be indicative of truth and deceit; an example is shown in Figure 5, wherein words marked in darker red appeared to be more important for the fake news detection task and vice versa. Although adding attention was not found to specifically improve detection performance, it provides some interpretability in the extracted features.

Limitation. The deep attention used is nice for providing insights into interpretability but is not seen to improve detection accuracy in Chen et al. [16] and thereby is more useful for qualitative studies on fake news rather than detection.

5.4.2 User Response Generation. Fake news detection research is generally split between content-based and feedback-based identification, i.e., detection using either article text or using secondary information from user engagements with the article. Qian et al. [89] provides a new approach to integrate the two sources of information directly targeted toward *early fake news detection*. User responses to articles can only be available after the news has propagated sufficiently to collect a sufficient number of user responses to apply detection methods. This can lead to increasing costs and is averse to the timeliness of detection. The article text, however, is available from the beginning. The early detection setting considered by Qian et al. [89] assumes that only the article text is available at the time of detection, which is similar to detection using content-based methods.

Approach. Unlike content-based methods, Qian et al. [89] propose a new strategy based on probabilistic generative modeling to still leverage rich secondary information without having access to responses for the article being classified. The strategy is to instead utilize historically collected articles and associated user responses to learn a probability distribution over user responses conditioned on the article, which can then be used to generate sampled responses to the article being classified. The conditional generative model (i.e., user response generator URG) is shown in

A fake news article sampled from the test set:			
... FDA quietly bans powerful life-saving intravenous Vitamin C ...It would be naive to think that the FDA endeavours to protect the public's health as its primary focus ...			
Top 20 words generated by URG in response:			
[[!]]	[[?]]	[[@]]	[[link]]
c	care	fda	food
false	health	intravenous	life
only	**problem**	protect	rich
tax	wait	watch	**why**

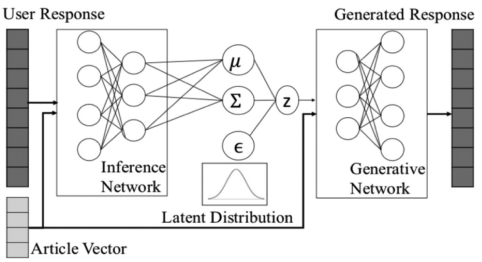


Fig. 6. Top 20 words from the generated response samples to a fake news article (left) and the user response generator network (right) [89]. Reprinted with permission from [89].

Figure 6. Article features are extracted using a convolutional neural network (CNN) trained on article word features. The obtained article representation along with the user response words are used to then train the URG. During detection, the article representation is first extracted using the CNN and used to condition the URG to generate sample responses to the article by sampling from the learned distribution. At this time, the URG receives only the article representation as an input, since there are no user responses collected at the time of detection, and relies on the learned distribution to generate sample responses. The final classification is done using the concatenated features of the extracted article representation from the CNN and the generated samples of user responses obtained from the URG conditioned on the article representation. The approach is found to have improved detection performance over using only the extracted article representation for detection.

Limitation. Qian et al. [89] only utilize a simplistic bag-of-words representation (vector of term frequency of words) of the textual features of the user response that are provided as input to the URG (generative model). Such representations might not capture more intricate syntactic and semantic information available in the word sequences of texts.

5.4.3 Stance Detection. User responses (comments) have also been utilized to improve detection performance by explicitly taking into account their stance toward the content being discussed. It is generally considered that each user response has one of four stances, namely support, deny, query, or comment. Recent works have considered multi-task learning approaches to jointly provide stance detection and veracity classification to improve classification accuracy by utilizing the interdependence in the two tasks [51, 64]. Both these works proposed multi-task learning architectures based on recurrent neural networks with shared and task-specific parameters.

Approach. Ma et al. [64] combine the tasks of stance detection and veracity classification where each task is assigned a shared GRU layer and a task-specific GRU layer. The purpose of the shared GRU layer is to capture patterns common to both tasks, whereas task-specific GRU enables the capture of patterns that are more important to one task than to the other. For instance, veracity classification relies more strongly on patterns directly conveying veracity such as "true" and "false," whereas patterns like "believe" and "don't think" can be more directed toward stance detection. The architecture proposed by Ma et al. [64] is depicted in Figure 7. The input is a sequence of posts (user responses) represented by text (tf-idf) features. The input sequence of vectors are converted to low-dimensional representations using a shared embedding layer and task-specific embedding layers and provided as input sequences to the respective shared and task-specific GRU layers. To enhance the interaction between the task-specific layer and shared layer, the hidden

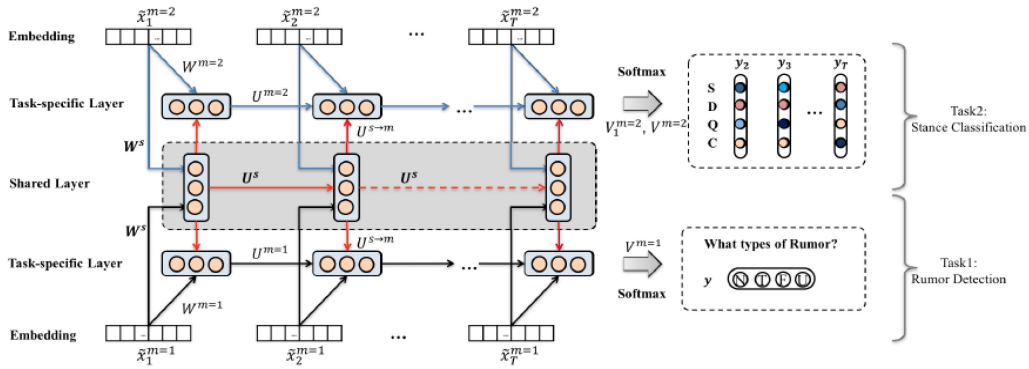


Fig. 7. Neural multi-task learning for stance detection and veracity classification jointly [64]. Reprinted with permission from [64].

state at each timestep of the task-specific GRUs are additionally made dependent on the hidden state from the shared GRU layer as shown in Figure 7. In the veracity classification task, the hidden state at the last timestep of the task-specific GRU is utilized for classification using a fully connected output layer with softmax activation specific to the veracity classification task. In the stance detection task, every post in the sequence is classified using a fully connected output layer with softmax activation specific to the stance detection task. Ma et al. [64] found that joint learning of stance detection and veracity classification tasks improves the performance of individual tasks. Utilizing shared and task-specific parameters were observed to be more beneficial than using only the shared parameters without the task-specific layer. Kochkina et al. [51] proposed a similar approach with shared and task-specific LSTM layers for rumor detection, stance classification, and rumor veracity classification jointly (rumor detection was specified as classifying the sequence of posts as responses to a rumor or non-rumor, and the rumor veracity classification task was specified as classifying the rumors as true, false, or remains unverified). In addition, each input sequence is considered as a sequence of posts along a particular branch of the propagation tree, instead of considering all posts by their temporal ordering regardless of the propagation tree as in Ma et al. [64].

Limitation. The process of annotating the stance of each user response, which is required as training data can be crowd-sourced but is still a time intensive process.

5.5 Response User Analysis

User reputation has long been a considered factor in many Internet services, such as *Amazon, LinkedIn, TripAdvisor*, and other e-commerce applications,[8] and on content-sharing platforms like Wikipedia, StackOverflow, and Quora. In the case of fake news propagation, users can be involved as sources or promoters of misinformation. In this section, we discuss efforts made toward characterization of users posting and propagating news over social media.

5.5.1 User Features. User features can be obtained based on two types of information. One is the features extracted from user profiles on social networks. The second is features extracted based on user behaviors from content sharing and response patterns of the user. For the first type of features, hand-crafted user features like registration age of users, number of followers, and the like are leveraged in Castillo et al. [15] along with other textual and propagation features for fake

[8] http://www.cnn.com/2000/TECH/computing/11/07/suing.ebay.idg/.

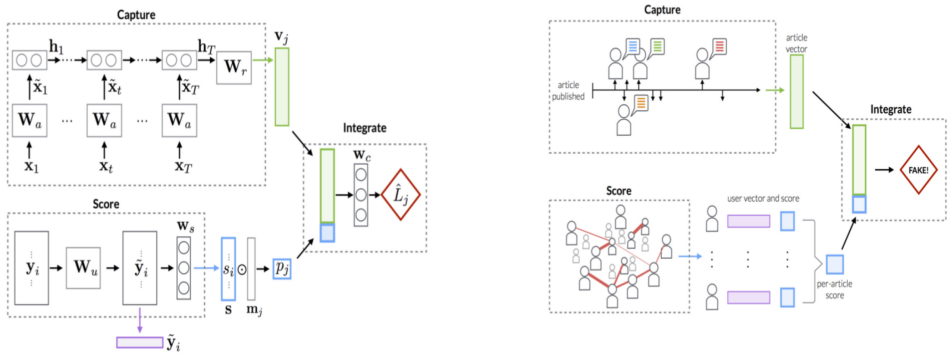


Fig. 8. Temporal pattern analysis and user group behavior analysis [96]. Reprinted with permission from [96].

news detection. Another way of obtaining user representations is proposed by Wu and Liu [121], who construct user representations using network embedding approaches on the social network graph. For the second type of features, Tacchini et al. [109] construct a feature vector consisting of the set of users who *liked* (responded) the source article or post and used these feature vectors to train a logistic regression model for classification of the source article or post as fake. The model essentially tries to capture patterns in user behaviors across articles for classification. Similarly, Qazvinian et al. [88] compute probability distributions over users involved in fake and true posts from collected user engagements to model user behavior patterns.

Limitation. Hand-engineered user features can be restrictive to one social media platform and not generalize to others. Individual user profiles and behaviors should be complemented by analysis of user group behaviors.

5.5.2 User Group Analysis. Most previous works treat individual user information separately, without considering group behaviors of users. Ruchansky et al. [96] noted that there are groups of users who share the same set of fake articles. According to the analytical study of information sharing on social media [25], "users tend to aggregate in communities of interest" and "users mostly tend to select and share content according to a specific narrative and to ignore the rest." Group behaviors of users are therefore beneficial for generalization of features across users and in identifying anomalous user groups responsible for the promotion of fake news.

Approach. To detect group behaviors, Ruchansky et al. [96] constructed a weighted user-user graph where an edge between users denotes the number of articles that both have co-engaged with. The user feature vector is constructed using singular value decomposition on the user-user graph, i.e., the user co-engagements matrix. The user features constructed this way are lower-dimensional distributed representation of users reflective of user group engagements and behaviors. The architecture of the detection model is shown in Figure 8. The capture component is used to extract temporal information from user engagements as explained in Section 5.3.2. The score component is used to learn a learn a suspiciousness score for each user and takes user feature vectors as input. It is observed that the method is able to generate meaningful suspiciousness scores over users, which are positively correlated with the engagements of the users in true and fake article cascades. In addition, the detection performance is observed to improve due to the integration the score component capturing user group behaviors into the overall architecture. Additionally, the benefit of the above method is that the user feature extraction does not require labeled

cascades, and we can extract user features from co-engagement information derived from unlabeled cascades itself. Labeled cascades are expensive to obtain.

Limitation. The main limitation is that in cases where new fake accounts and social bots are created specifically to spread fake news, there might be no collected user behaviors for such accounts and no direct way to generalize extracted user features to these accounts.

Limitations of existing feedback-based methods. One significant drawback of existing feedback-based classification methods discussed above is that the models are trained on a snapshot of the user responses generally collected after or toward the end of the propagation process, when sufficient responses are available. This explains a drop in performance on early detection using the trained models when fewer responses are available. The methods do not have the ability to update their state based on incrementally available user responses.

6 INTERVENTION-BASED SOLUTIONS

The scope of the methods discussed in content-based and feedback-based identification is limited to classifying news from a snapshot of features extracted from social media. For practical applications, we need techniques that can dynamically interpret and update the choice of actions for combating fake news based on real-time content propagation dynamics. We discuss the works that provide such computational methods and algorithms in this section.

6.1 Mitigation Strategies

Exposure to fake news can lead to a massive shift in public opinion as noted in Section 1. We discuss several strategies aimed at reversing this effect by strategically introducing true news into the social media platforms such that users can be exposed to the truth and the impacts of fake news on user opinion can be mitigated. Computational methods designed for this purpose require us to first consider some of the widely used information diffusion models. There are several well-known models used to represent information diffusion in social networks, such as the *Independent Cascade* (IC) and *Linear Threshold* (LT) model [47] and point process models such as *Hawkes Process* model [129]. The IC model represents each edge with a parameter $p_{u,v}$ denoting the strength of influence of user u on user v. The diffusion process starts with a set of seed nodes assumed to be activated at the first timestep. At each timestep of the diffusion process, a node u activated at timestep t, independently makes a single activation attempt on each inactive neighbor v. The activation succeeds with probability $p_{u,v}$ and a node once activated remains activated throughout the diffusion process. In the LT model the edge parameters represent a weight of influence and additionally each user has a separate (uniformly and independently) random threshold parameter and the sum of weights of incoming edges should be less than 1. Accordingly, activation of a user occurs if the weights of all its activated neighbors exceed its threshold at a given timestep of the diffusion process. Point process models are defined differently from the above two by defining an intensity function that we discuss later in Section 6.1.3.

6.1.1 Decontamination. Nguyen et al. [75] proposed a strategy to decontaminate users that have been exposed to fake news. The diffusion process is modeled using the *Independent Cascade* or *Linear Threshold* model [47] described earlier. A greedy algorithm is designed to select the best set of seed users from which to start the diffusion process for true news, so that at least a β-fraction of the users can be decontaminated. The algorithm is a simple greedy algorithm that iteratively selects the next best user to include into the seed set based on the marginal gains obtained by the inclusion of the user (i.e., the number of users that will be activated or reached by the true news in expectation, if the seed set did additionally include the chosen user). The iterative selection of

seed users is continued till the objective of decontaminating β-fraction of the contaminated users in expectation can be achieved. Nguyen et al. [75] show that this greedy algorithm has a $(1 - 1/e)$ approximation ratio with respect to the optimal seed set that should be selected.

Limitation. One drawback of the proposed approach is that it is suggested as a corrective measure *after* the spread of fake news. Second, the number of seeds is not fixed, and if the damage done is large, then the cost of correction by activating or incentivizing more seed users to start the true news cascade might be prohibitive.

6.1.2 Competing Cascades. Several other works propose a more interesting intervention strategy based on competing cascades, wherein a true news cascade is introduced to compete with the fake news cascade as the fake news originates and begins to propagate through the network rather than *after* its propagation. In particular, Budak et al. [11] and He et al. [39] formulated an *influence blocking maximization* objective as finding an optimal strategy to disseminate true news in the presence of a misinformation cascade by strategically selecting k seed users, with the objective of minimizing the number of users who at the end of the diffusion are activated by the fake news campaign instead of the true news one. The model assumes that a user once activated by either the fake or true cascade remains activated under that cascade. Each edge contains a separate set of diffusion model parameters, one for true news and the other for fake. It is reasonable to model the two types of diffusion processes separately, as the strength of influence under true news might not be the same when sharing fake news. Budak et al. [11] showed that a similar greedy algorithm as discussed in Section 6.1.1 achieves a $(1 - 1/e)$ approximation ratio to the optimal under the restriction that the true news influence probabilities are all 1 or exactly equal to the negative influence probabilities. He et al. [39] similarly showed that under the *Linear Threshold* model, the greedy algorithm ratio is $(1 - 1/e)$ without requiring restrictions on the influence weights.

Limitation. The selection of seed users for the true news cascade happens once at the beginning of the process in response to a detected fake news cascade, after which both cascades continue to propagate competitively without external moderation. Second, a user being "activated" represents both exposure and re-sharing, since the model does not differentiate between them and assumes that an activated user is exposed and always attempts to activate its neighbors through forced re-sharing.

6.1.3 Multi-stage Intervention. Extension to a multi-stage intervention strategy was proposed by Farajtabar et al. [27] to allow external interventions to adapt as necessary to the observed propagation dynamics of fake news. Specifically, Farajtabar et al. [27] considered a different model of social influence based on multivariate point processes wherein past news-sharing events trigger future news-sharing events based on the intensities of influence and time delay between events.

Approach. In multivariate point processes, each user i's (news-sharing) events are triggered by their base intensity μ_i and past (news-sharing) events with some decay function g that depends on the time gap between the current time and the past event and the past events of other users j, whose influence is captured by $\alpha_{i,j}$. The intensity function of user i is then $\lambda_i(t) = \mu_i + \sum_{t_j < t} \alpha_{i,j} g(t - t_j)$ and the probability of an event conditioned on past events is proportional to the intensity function. In [27], the authors find that this model can naturally capture the diffusion of fake (true) news and allow one to separately capture being "exposed" to fake (true) news vs. being exposed and "sharing" fake (true) news, addressing the drawbacks of works mentioned in Section 6.1.2 that cannot make that distinction and assume that if a user is exposed to fake (true) news, then the user also shares the fake (true) news with his/her social media connections. Farajtabar et al. [27] model fake news with one point process $F(t)$ and its counteracting true (mitigation) news with

another $M(t)$. $M(t)$ represents a vector with the number of times each user shares a mitigation event. $\mathcal{M}(t) = BM(t)$, where B is the adjacency matrix represent event exposures, implying that a user is exposed whenever she or her neighbor shares the news. Similarly, fake news event shares and exposures counts are represented with $F(t)$ and $\mathcal{F}(t)$. The purpose of the intervention at any stage is to externally incentivize certain users (social network moderators can recommend, display, or share the true event with them directly) to enable increased sharing of true news over the network that can counteract the fake news process. At each stage of the intervention certain budget and user activity constraints are imposed. The model allows us to solve for the optimal amount of external incentivization needed on every user at each stage under imposed constraints that will be enough to achieve a desired objective/reward from the intervention (such as minimizing the difference between fake and true news exposures). Farajtabar et al. [27] provide a reinforcement learning-based policy iteration framework for optimization of the proposed multi-stage intervention objective to derive the optimal amount of external incentivization.

Limitation. The method assumes that one has already identified fake news and is tracking its propagation through the network that is not trivial. Moreover, it could be arguable that once we have identified the fake news, we should simply use direct recommendations of true news to the users who we know are exposed to the fake news and remove the fake news content from the platform to prevent future spread.

6.2 Identification Strategies

As stated in Section 1, studies on information cascades reveal that information is readily and rapidly transmitted over the network, even when it is of dubious veracity [31]. Given that even fake news can propagate exponentially quickly through reshares on social media, intervention efforts directed toward early identification and containment of fake news are of primary significance. In this section, we discuss works that suggest different mechanisms and intervention efforts to *actively* detect and prevent the spread of fake news on social media.

6.2.1 Network Monitoring. Intervention strategy based on network monitoring involves intercepting information from a list of suspected sources of fake news using computer-aided social media accounts or real paid user accounts, whose role is to filter the information they receive and block what they consider to be fake news. The network monitor placement can be determined by finding the cut or partition of the network that has the highest probability of transmission across the cut and a maximum of k users on the side of suspected sources [3]. Another possible network monitoring placement solution is based on a Stackelberg game between attacker and defender nodes [125]. The solution to the game allows the network administrator to select the set of nodes to monitor. Multiple monitoring sites with heterogeneous detection capabilities in both works are motivated by two ideas. One that having multiple check-points where fake news can be detected reduces the chance that the fake news would go undetected. Second, having multiple human or machine classifiers improves detection robustness, because something that is missed by one fact-checker might be captured by another.

Limitation. In changing network topologies, network monitoring solutions would require to update their strategy based on the change. Also, network monitoring on large-scale networks could be expensive due to the size of the network.

6.2.2 Crowd-sourcing. Another set of intervention efforts are designed to leverage crowd-sourcing mechanisms introduced on social media platforms that allow users to report or flag fake news articles. Instead of explicit network monitors as used by the approaches discussed earlier, these works implicitly monitor the network using crowd-sourced user feedback. However, a major

drawback of the reporting or flagging mechanism is its possible misuse by adversaries. In addition, even signals from trustworthy users can be noisy, since users are not all equally good at identifying misinformation. A data-driven assessment by Freeman [30] on data from LinkedIn suggests that only 1.3% of the users show measurable and repeatable skills at reporting actions performed by fake accounts.

Approach. One way to leverage crowd-sourced signals to prioritize fact checking of news articles was proposed in Kim et al. [48] by capturing the tradeoff between the collection of evidence (flags) vs. the harm caused from more users being exposed to fake news (exposures) to determine when the news needs to be verified. They modeled the fact-checking process and events using point process models and derived the optimal intensity of fact checking that is proportional to the rate of exposures to misinformation and the collected evidence as flags. To more accurately leverage user flags, Tschiatschek et al. [112] designed an online learning algorithm to jointly infer the flagging accuracies of users while identifying fake news. The algorithm selects k news to inspect at each stage by greedily picking the ones that maximize the total expected utility. The total expected utility is an aggregation of the expected utility (reward) over each stage t of the intervention, starting from stage 1 to the last stage T (that the algorithm is allowed to fix), and is mathematically defined as follows:

$$\text{Util}(T, \text{ALGO}) = \sum_{t=1}^{T} \mathbb{E}\left[\sum_{s \in S^t} 1_{\{y^\star(s)=f\}} \text{val}^t(s)\right],$$

where $\text{val}^t(s)$ is the expected reward at stage t from the inspection and debunked of fake news from the news selected for fact checking at stage t. The reward at stage t can be modeled as the number of users prevented from seeing the fake news because of the intervention and debunking at stage t. Tschiatschek et al. [112] note that this quantity can be estimated by modeling and simulating the future spread to get the estimated number of users prevented from being exposed to the fake news that was debunked. $y^\star$ is the label of the article, estimated using Bayesian inference from $(\theta_{u,f}, \theta_{u,\bar{f}})$, which are parameters used to model the flagging accuracy of a user and are estimated based on the history D of fact-checked news by sampling from the posterior distributions $P(\theta_{u,f}|D)$ and $P(\theta_{u,\bar{f}}|D)$.

Limitation. Crowd-sourcing methods that essentially utilize reporting mechanisms trade off the number of users affected (exposed to fake news) with the number of reports (flags) needed to make a reliable prediction.

6.2.3 User Behavior Modeling. Apart from solutions based on explicit or implicit network monitoring, another proposed solution is based on a model of user behavior and news sharing [79]. The proposed model is developed to determine how to prioritize fact checking of news articles by determining the optimal time of inspection of an article based on the proposed model of news sharing. The sharing of news is modeled as a sequential process among rational agents, wherein each agent chooses whether to verify the article and whether to share it with the next agent. By rational agents, it is assumed that agents intend to share only true news. Under this model, the intervention can be designed as an optimal stopping problem solved using dynamic programming to find the optimal time of inspection. The model characterizes agent judgments in verifying fake news as well as the sequential interaction between agents, which are useful in determining the optimal time and priority of inspection. For instance, if the article is likely to be true or if is believed to be fake but more likely to be verified by the agents themselves before sharing, then the need for inspection is lowered.

Limitation. A drawback of the proposed method is that it is restricted to "tree" structured networks and cannot generalize to complex network topologies.

Limitations of existing intervention-based methods. Intervention-based methods are more difficult to evaluate and test out, especially in complex environments like this, with lots of interdependent interactions. Also, they might make restrictive assumptions in certain cases that could limit their applicability.

7 EXISTING DATASETS

The development of novel solutions to fake news detection has often been limited by data quality. To facilitate future research on this topic, we have compiled a comprehensive list of datasets related to the fake news detection task. Due to different research purposes, the data collections could vary significantly. For example, some datasets focus solely on political statements while others consist of open-domain news/articles. In addition, datasets could vary depending on what labels are provided, how are the labels collected, what types of text contents are included (e.g., source claims, user responses, etc.), whether temporal and propagation information is recorded, and the size of the collection. Therefore, we summarize these properties of the existing datasets.

7.1 Datasets Enumeration

The following is a comprehensive list of existing datasets. For named datasets, we use existing names as is; for unnamed ones, we assign suitable names for referencing.

7.1.1 LIAR. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection [118] (https://www.cs.ucsb.edu/william/data/liar_dataset.zip).

7.1.2 Twitter. Detecting Rumors from Microblogs with Recurrent Neural Networks [61] (http://alt.qcri.org/~wgao/data/rumdect.zip).

7.1.3 Weibo. Detecting Rumors from Microblogs with Recurrent Neural Networks [61] (http://alt.qcri.org/~wgao/data/rumdect.zip).

7.1.4 FacebookHoax. Some Like it Hoax: Automated Fake News Detection in Social Networks [109] (https://github.com/gabll/some-like-it-hoax/tree/master/dataset).

7.1.5 PHEME-R. Analyzing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads [133] (https://figshare.com/articles/PHEME_rumour_scheme_dataset_journalism_use_case/2068650).

7.1.6 PHEME. All-in-one: Multi-task Learning for Rumour Verification [52] (https://figshare.com/articles/PHEME_dataset_for_Rumour_Detection_and_Veracity_Classification/6392078).

7.1.7 Cred-1. Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media [85] (http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/impact/web-credibility-analysis/).

7.1.8 Cred-2. Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media [85] (http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/impact/web-credibility-analysis/).

7.1.9 FakevsSatire. Fake News vs. Satire: A Dataset and Analysis [35] (https://github.com/jgolbeck/fakenews).

7.1.10 BuzzfeedNews. Hyperpartisan Facebook Pages Are Publishing False And Misleading Information At An Alarming Rate [103] (https://github.com/BuzzFeedNews/2016-10-facebook-fact-check)/

7.1.11 KaggleFN. Kaggle fake news dataset (https://www.kaggle.com/mrisdal/fake-news).

7.1.12 NewsFN. (GeorgeMcIntire/fake_real_news_dataset https://github.com/GeorgeMcIntire/fake_real_news_dataset).

7.1.13 BuzzfeedPolitical. This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News [43] (https://github.com/BenjaminDHorne/fakenewsdata1).

7.1.14 Political-1. This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News [43] (https://github.com/BenjaminDHorne/fakenewsdata1).

7.1.15 KaggleEmergent. Kaggle rumors dataset (https://www.kaggle.com/arminehn/rumor-citation).

7.1.16 NewsFN-2014. Fact Checking: Task definition and dataset construction [115] (https://sites.google.com/site/andreasvlachos/resources/FactChecking_LTCSS2014_release.tsv?attredirects=0).

7.1.17 NewsTrustData. Leveraging Joint Interactions for Credibility Analysis in News Communities [71] (https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/impact/credibilityanalysis/).

7.1.18 FakeNewsNet-1. FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media [100] (https://github.com/KaiDMML/FakeNewsNet)

7.1.19 FakeNewsNet-2. FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media [100] (https://github.com/KaiDMML/FakeNewsNet)

7.1.20 Twitter15. Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning [63] (https://www.dropbox.com/s/7ewzdrbelpmrnxu/rumdetect2017.zip?dl=0).

7.1.21 Twitter16. Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning [63] (https://www.dropbox.com/s/7ewzdrbelpmrnxu/rumdetect2017.zip?dl=0).

7.1.22 FEVER. FEVER: A Large-scale Dataset for Fact Extraction and VERification [111] (http://fever.ai/data.html).

7.1.23 FNC-1. Fake News Challenge: Stance Detection Dataset (https://github.com/FakeNewsChallenge/fnc-1).

7.2 Dataset Characteristics

We provide a summarization of the characteristic features of listed datasets in Tables 2, 3, and 4 based on the task and application context, information content features, and collected user responses features.

Table 2. Datasets Characteristics

Dataset	Task/Application	Task labels	Content type	Annotator
LIAR	fake news detection	pants-fire, false, barely true, half-true, mostly true, and true[1]	political statements	PolitiFact
FakevsSatire	fake news detection	fake, satire[1]	political news	researchers
NewsFN	fake news detection	fake, real	news articles	unspecified
BuzzfeedPolitical	fake news detection	fake, real	political news	verified list
Political-1	fake news detection	fake, real, satire	political news	Zimdars websites list
NewsFN-2014	fake news detection	true, mostly true, half true, mostly false, false[1]	fact-checked claims	Channel4 and PolitiFact
NewsTrustData	credibility assessment	qualitative scores[11]	news articles	NewsTrust member
Twitter	rumor classification	rumor, non-rumor	fact-checked claims	mapped from Snopes[5]
Weibo	rumor classification	rumor, non-rumor	fact-checked claims	Sina community management center
Twitter15	rumor classification	rumor (false, true, remains unverified), non-rumor	fact-checked claims	mapped from Snopes[5]
Twitter16	rumor classification	rumor (false, true, remains unverified), non-rumor	fact-checked claims	mapped from Snopes[5]
PolitiFact	fake news detection	fake, real	political statements	PolitiFact
GossipCop	fake news detection	fake, real	entertainment news	GossipCop
FacebookHoax	hoax detection	hoax, non-hoax	sicentific, conspiracy	non-hoax as scientific
PHEME-R[12]	rumor analysis	rumor only (false, true, remains unverified)[2]	newsworthy stories	journalists,[6] crowd-sourcing[7]
PHEME	rumor classification	rumor (false, true, remains unverified), non-rumor	newsworthy stories	journalists[6], crowd-sourcing[7]
BuzzfeedNews	fake news detection	mostly true, mixture, mostly false, no factual content	political news	unspecfied
KaggleEmergent	rumor classification	rumor (false, true, unverified)	fact-checked claims	Emergent
KaggleFN	fake news detection	only fake	news articles	BS Detector[8]
Cred-1	fact extraction	false, true	fact-checked claims	mapped from Snopes[9]
Cred-2	fact extraction	only false	wikipedia hoax	verified list[9]
FEVER	fact extraction	supported, refuted, not enough info[4]	constructed claims[3]	trained annotators
FNC-1	stance detection	agrees, disagrees, discusses, unrelated	news articles	FakeNewsChallenge[10]

Task/Application and Annotation. [1]with justifications [2]plus posts stance, certainty, evidentiality [3]using Wikipedia [4]with evidence [5]mapping policy not specified [6]for class label [7]other annotations [8]not fact checked by humans [9]also provides Search Engine results for claims (but without annotating stance) [10]stance of the body toward the title [11](objectivity, correctness, bias, credibility) [12]used in RumorEval Task.

Table 3. Dataset Characteristics

Dataset	Total claims	Collection period	Number of claims (split by label)
LIAR	12.8K	2007–2016	roughly equal
FakevsSatire	486	2016–2017	58% fake
NewsFN	6,331		roughly equal
BuzzfeedPolitical	120	Jan.–Oct. 2016	roughly equal
Political-1	225		roughly equal
NewsFN-2014	221	2013–2014	68% half true, 50% false, rest ~35%
NewsTrustData	82K	per article	not applicable to qualitative scores
Twitter	992	Mar.–Dec. 2015	roughly equal
Weibo	4,664		roughly equal
Twitter15	1,490		roughly equal
Twitter16	818		roughly equal
PolitiFact	488	dynamic	roughly equal
GossipCop	3570	dynamic	19% fake
FacebookHoax	15.5K		60% hoax
PHEME-R	330		50% true, 20% false, 30% unverified
PHEME	6,425		60% non-rumor, rumor: 16% true, 10% false, 10% unverified
BuzzfeedNews	2,282	19–27 Sep 2016	38% mostly false / mixture
KaggleEmergent	2,145	Jan.–Dec. 2017	26% false, 34% true, 40% unverified
KaggleFN	13K	Oct.–Nov. 2016	all fake
Cred-1	4,856		74% false
Cred-2	157		all false
FEVER	185K		55% support, 20% refute, 25% other
FNC-1	50K		7% agree, 2% disagree, 18% discuss[5]

Information content to be classified (claims, articles, etc.).

Table 4. Dataset Characteristics

Dataset	Avg/claim	Time-stamp	Text	User info	Tree	Network	Platform
Twitter	1,111	y	y	y			Twitter
Weibo	816	y	y	y			Weibo
Twitter15	223	y	y	y	y		Twitter
Twitter16	251	y	y	y	y		Twitter
PolitiFact	357	y	y	y		y	Twitter
GossipCop	10	y	y	y		y	Twitter
FacebookHoax	156[1]	y	y	y			Facebook
PHEME-R	15	y	y	y	y	y	Twitter
PHEME	16	y	y	y	y		Twitter
BuzzfeedNews	9,792[2]						Facebook
KaggleEmergent	7,187[3]						not specified
KaggleFN	27[2]						Facebook

Collected responses (user engagements) on social media. [y=yes, blank=no]. [1]only likes [2]only counts (shares, reactions, comments) [3]only counts (shares).

7.2.1 Task and Application Context. In Table 2, we mention the task/application context for each dataset along with the task labels provided in the dataset. We additionally mention the content types (domains) such as political news, entertainment news, Wikipedia statements, and so on, that are encompassed by contents of each dataset. Last, we provide the annotation scheme used in the dataset that provides crucial insight into the data quality. For instance, the KaggleFN dataset consists of news articles that are marked as fake by the `BS-Detector` application and not through journalistic or human verification, which means that the dataset labels could be noisy and, moreover, training a model to learn those labels is essentially training a model to mimic `BS-Detector` [100].

7.2.2 Information Content. In Table 3, we provide additional details about the information content that are made available in each dataset. The information content (claim) to be verified can be an article, post, political statement, news stories, and so on, as mentioned under Content Types in Table 2. Here we mention the total number of claims in each dataset. Total claims is the number of information contents that need to be verified. In addition, we provide the division of claims by task labels such as the percentage of fake vs. true articles in the dataset, which are useful for accessing applicability of machine-learning models to a specific dataset. Last, we provide the collection period, that is, the period in which the collected articles/claims were published or posted, which, much like the domain (content types), is an important characteristic of the dataset. When the collection period is left blank, it means that it is unspecified in the dataset.

7.2.3 User Responses. Last, in Table 4 we provide features of user responses collected from user engagements with the information content on social media. Note that this information is only provided in a subset of the datasets. We mention different features of the user responses in the table such as average responses per claim, which is a popularly mentioned statistic in existing datasets. However, there can be high variation in the number of responses, based on virality of collected stories, and the platform used for collection such as Twitter, Weibo. The platform is also mentioned in the table for each dataset.[9] In addition, we note whether timestamp information, text information (in comments/replies), user information (user participating in the response), propagation tree information (who replies to whom), and social network information of users (follower-followee relationships) is provided in each dataset. Together, the three tables provide comprehensive insights into all existing datasets, facilitating a comparative analysis between them, to identify the potential shortcomings of existing datasets, as well as to identify the best choice of a dataset for a particular application/task or to determine its suitability for training a specific detection model.

8 CONCLUSION AND FUTURE WORK

The literature surveyed here has demonstrated significant advances in addressing the identification and mitigation of fake news. Nevertheless, there remain many challenges to overcome in practice. We now outline three concrete directions for future research that can further advance the landscape of computational solutions.

(1) **Dynamic knowledge bases.** Although we can design many features that can potentially help us determine the truthfulness of a news article, the truthfulness is still ultimately defined by the statements it makes. The greatest challenge in developing automated

[9]In accordance with the platform Twitter's information-sharing policy, public release and distribution of tweet contents and metadata are restricted. Twitter-based datasets generally release tweet ids and user ids from which complete information about contents and metadata is obtainable through Twitter Search API. For these datasets, for information that is obtainable from Twitter, we mark the information as available in our tables regardless of whether it is released in the dataset directly or through the Twitter id information.

fact-checking methods is the construction of dynamic knowledge bases, which can be regularly and automatically updated to reflect the changes occurring in a fast-paced world.

(2) **New intervention strategies.** For the design of useful intervention strategies for different environments, the most important question that needs to be answered is which environmental factors are most conducive to the spread of fake news. Studying and characterizing the relationship between user actions and utilities at the microscopic level of the individual, and the macroscopic impact in different networked environments, will be essential for explaining the spread of fake news and finding the best intervention strategies suited to that environment. Another interesting direction is toward *educational* interventions. Recently, Gillani et al. [34] studied the effect of educating people about the polarization of their social ego-networks using visualization tools that show users their social connections along with the inferred political orientations of other users in their network to determine its impact on the user's connection diversity after the treatment. Another study proposed an *active inoculation* strategy using a "fake news game," where participants were asked to create content by using misleading tactics from the perspective of fake news creators [94]. The authors found preliminary evidence of its effectiveness in reducing perceived reliability and persuasiveness of fake news articles in a randomized field study of 95 high school students.

(3) **Datasets for intent detection.** Current datasets generally provide binary labels of information as fake or true. However, a more fine-grained classification of information by intent might be especially beneficial in identifying *truly* fake news from closely related information such as satire and opinion news. In the list of fake websites maintained in [131], there is a *type* label allowing up to three types for each website with tags such as political, satire, or biased. Some recent works have considered classification of fake vs. satire news [35] and fake vs. hyperpartisan news [87], and we believe that this is an important direction for the future.

A APPENDIX

We consolidate quantitative results on fake news (rumor) detection in terms of classification accuracy for several of the methods discussed earlier on a representative sample of the datasets namely Twitter collected in [89], Weibo and Twitter collected in [61], Twitter15 and Twitter16 collected in [63]; based on experimental evaluation performed in several works [61, 65, 89, 96]. While the list of methods and datasets is not exhaustive, there are other methods such as Ma et al. [64], Kochkina et al. [51] and other datasets such as PHEME, PolitiFact, GossipCop [51, 100] that could be evaluated, we regard a more extensive bench-marking based on specific application contexts with datasets of varying content types and collection periods as a direction for future work. Tables 5 and 6 contain the consolidated quantitative results of classification accuracy for detection under content-based and feedback-based methods.

Table 5. Content-based (w/o User Responses at Test Time) Methods

Methods	Weibo [61]	Twitter [89]
Ott et al. [77] - LIWC	66.06	62.13
Ott et al. [77] - POS	74.77	70.34
Ott et al. [77] - n-gram	84.76	80.69
Wang [118]	86.23	83.24
Qian et al. [89]	89.84	88.83

Classification accuracy (%). Results consolidated from [89].

Table 6. Feedback-based Methods

Methods	Weibo [61]	Twitter [61]	Twitter15 [63]	Twitter16 [63]
Zhao et al. [128]	73.2	64.4	40.9	41.4
Castillo et al. [15]	83.1	73.1	45.4	46.5
Kwon et al. [56]	84.9	77.2	56.5	58.5
Ma et al. [62]	85.7	80.8	54.4	57.4
Ma et al. [61]	91.1	88.1	64.1	63.3
Ruchansky et al. [96]	95.3	89.2	-	-
Wu et al. [120]	x	x	49.3	51.1
Ma et al. [63]	x	x	66.7	66.2
Ma et al. [65]	x	x	72.3	73.7

Feature engineering, temporal, propagation. Classification accuracy (%). Results consolidated from [65, 96].
x = Cannot be evaluated as per dataset features (information unavailable for method).

ACKNOWLEDGMENTS

We thank the reviewers and moderators for their invaluable comments and inputs on earlier versions of this manuscript.

REFERENCES

[1] Mohammed Ali and Timothy Levine. 2008. The language of truthful and deceptive denials and confessions. *Commun. Rep.* 21, 2 (2008), 82–91.
[2] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *J. Econ. Perspect.* 31, 2 (2017), 211–36.
[3] Marco Amoruso, Daniele Anello, Vincenzo Auletta, and Diodato Ferraioli. 2017. Contrasting the spread of misinformation in online social networks. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1323–1331.
[4] Jessikka Aro. 2016. The cyberspace war: Propaganda and trolling as warfare tools. *Eur. View* 15, 1 (2016), 121–132.
[5] Michael Barthel, Amy Mitchell, and Jesse Holcomb. 2016. Many Americans believe fake news is sowing confusion. *Pew Res. Center* 15 (2016), 12.
[6] Alessandro Bessi and Emilio Ferrara. 2016. Social bots distort the 2016 US presidential election online discussion. *First Monday* 21, 11 (2016).
[7] Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, Edward Finegan, and Randolph Quirk. 1999. *Longman Grammar of Spoken and Written English*. Vol. 2. MIT Press Cambridge, MA.
[8] Phil Blunsom, Edward Grefenstette, and Nal Kalchbrenner. 2014. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.
[9] Glynis Bogaard, Ewout H. Meijer, Aldert Vrij, and Harald Merckelbach. 2016. Scientific content analysis (SCAN) cannot distinguish between truthful and fabricated accounts of a negative Event. *Front. Psychol.* 7 (2016), 243.
[10] Gary D. Bond and Adrienne Y. Lee. 2005. Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language. *Appl. Cogn. Psychol.* 19, 3 (2005), 313–329.
[11] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. 2011. Limiting the spread of misinformation in social networks. In *Proceedings of the 20th International Conference on World Wide Web*. ACM, 665–674.
[12] David B. Buller and Judee K. Burgoon. 1996. Interpersonal deception theory. *Commun. Theory* 6, 3 (1996), 203–242.
[13] David B. Buller, Judee K. Burgoon, Aileen Buslig, and James Roiger. 1996. Testing interpersonal deception theory: The language of interpersonal deception. *Commun. Theory* 6, 3 (1996), 268–288.
[14] Carlos Carvalho, Nicholas Klagge, and Emanuel Moench. 2011. The persistent effects of a false news shock. *J. Empir. Finance* 18, 4 (2011), 597–615.
[15] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*. ACM, 675–684.

[16] Tong Chen, Xue Li, Hongzhi Yin, and Jun Zhang. 2018. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 40–52.

[17] Yimin Chen, Niall J. Conroy, and Victoria L. Rubin. 2015. Misleading online content: Recognizing clickbait as false news. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*. ACM, 15–19.

[18] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. Association for Computational Linguistics, Doha, Qatar, 1724–1734. https://doi.org/10.3115/v1/D14-1179

[19] Sahil Chopra, Saachi Jain, and John Merriman Sholar. 2017. Towards Automatic Identification of Fake News: Headline-Article Stance Detection with LSTM Attention Models. Retrieved on 2019 from https://johnsholar.com/pdf/CS224NPaper.pdf.

[20] Alton Y. K. Chua, Sin-Mei Cheah, Dion Hoe-Lian Goh, and Ee-peng Lim. 2016. Collective rumor correction on the death hoax of a political figure in social media. In *Proceedings of thePacific Asia Conference on Information Systems*.

[21] Michael Collins and Nigel Duffy. 2002. Convolution kernels for natural language. In *Advances in Neural Information Processing Systems*. 625–632.

[22] Nicole A. Cooke. 2017. Posttruth, truthiness, and alternative facts: Information behavior and critical information consumption for a new age. *Libr. Quart.* 87, 3 (2017), 211–221.

[23] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2015. Fame for sale: Efficient detection of fake Twitter followers. *Decis. Supp. Syst.* 80 (2015), 56–71.

[24] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee, 273–274.

[25] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proc. Natl. Acad. Sci. U.S.A.* 113, 3 (2016), 554–559.

[26] Lawrence N. Driscoll. 1994. A validity assessment of written statements from suspects in criminal investigations using the scan technique. *Police Stud.: Int'l Rev. Police Dev.* 17 (1994), 77.

[27] Mehrdad Farajtabar, Jiachen Yang, Xiaojing Ye, Huan Xu, Rakshit Trivedi, Elias Khalil, Shuang Li, Le Song, and Hongyuan Zha. 2017. Fake news mitigation via point process based intervention. In *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70. PMLR, 1097–1106.

[28] Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics, 171–175.

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

[30] David Mandell Freeman. 2017. Can you spot the fakes?: On the limitations of user feedback in online social networks. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1093–1102.

[31] Adrien Friggeri, Lada A. Adamic, Dean Eckles, and Justin Cheng. 2014. Rumor cascades. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM'14)*.

[32] Christie M. Fuller, David P. Biros, and Rick L. Wilson. 2009. Decision support for determining veracity via linguistic-based cues. *Dec. Supp. Syst.* 46, 3 (2009), 695–703.

[33] Kiran Garimella, Aristides Gionis, Nikos Parotsidis, and Nikolaj Tatti. 2017. Balancing information exposure in social networks. In *Advances in Neural Information Processing Systems*. 4663–4671.

[34] Nabeel Gillani, Ann Yuan, Martin Saveski, Soroush Vosoughi, and Deb Roy. 2018. Me, my echo chamber, and I: Introspection on social media polarization. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 823–831.

[35] Jennifer Golbeck, Matthew Mauriello, Brooke Auxier, Keval H. Bhanushali, Christopher Bonk, Mohamed Amine Bouzaghrane, Cody Buntain, Riya Chanduka, Paul Cheakalos, Jennine B. Everett, et al. 2018. Fake news vs satire: A dataset and analysis. In *Proceedings of the 10th ACM Conference on Web Science*. ACM, 17–21.

[36] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. 2014. Tweetcred: Real-time credibility assessment of content on twitter. In *Proceedings of the International Conference on Social Informatics*. Springer, 228–243.

[37] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. 2013. Faking sandy: Characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 729–736.

[38] L. Harriss and K. Raymer. 2017. Online Information and Fake News.
[39] Xinran He, Guojie Song, Wei Chen, and Qingye Jiang. 2012. Influence blocking maximization in social networks under the competitive linear threshold model. In *Proceedings of the 2012 SIAM International Conference on Data Mining*. SIAM, 463–474.
[40] Alfred Hermida. 2012. Tweets and truth: Journalism as a discipline of collaborative verification. *J. Pract.* 6, 5–6 (2012), 659–668.
[41] Kathleen Higgins. 2016. Post-truth: A guide for the perplexed. *Nat. News* 540, 7631 (2016), 9.
[42] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neur. Comput.* 9, 8 (1997), 1735–1780.
[43] Benjamin D. Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *arXiv preprint arXiv:1703.09398* (2017).
[44] Fang Jin, Edward Dougherty, Parang Saraf, Yang Cao, and Naren Ramakrishnan. 2013. Epidemiological modeling of news and rumors on twitter. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis*. ACM, 8.
[45] Mark Johnson. 1998. PCFG models of linguistic tree representations. *Comput. Ling.* 24, 4 (1998), 613–632.
[46] U Kang, Hanghang Tong, and Jimeng Sun. 2012. Fast random walk graph kernel. In *Proceedings of the 2012 SIAM International Conference on Data Mining*. SIAM, 828–838.
[47] David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 137–146.
[48] Jooyeon Kim, Behzad Tabibian, Alice Oh, Bernhard Schölkopf, and Manuel Gomez Rodriguez. 2018. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining (WSDM'18)*.
[49] Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. Association for Computational Linguistics, Doha, Qatar, 1746–1751. https://doi.org/10.3115/v1/D14-1181
[50] Masahiro Kimura, Kazumi Saito, and Hiroshi Motoda. 2009. Efficient estimation of influence functions for SIS model on social networks. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'09)*. 2046–2051.
[51] Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task Learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, 3402–3413. http://aclweb.org/anthology/C18-1288.
[52] Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task Learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, 3402–3413. http://aclweb.org/anthology/C18-1288.
[53] Shimon Kogan, Tobias J. Moskowitz, and Marina Niessner. 2018. Fake news: Evidence from financial markets. *SSRN: 3237763* (2018). https://doi.org/10.2139/ssrn.3237763
[54] Srijan Kumar and Neil Shah. 2018. False information on web and social media: A survey. *arXiv preprint arXiv: 1804.08559* (2018).
[55] Srijan Kumar, Robert West, and Jure Leskovec. 2016. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 591–602.
[56] Sejeong Kwon, Meeyoung Cha, and Kyomin Jung. 2017. Rumor detection over varying time windows. *PloS ONE* 12, 1 (2017), e0168344.
[57] John Michael Lake. 2014. Fake Web Addresses and Hyperlinks. US Patent 8,799,465.
[58] David F. Larcker and Anastasia A. Zakolyukina. 2012. Detecting deceptive discussions in conference calls. *J. Account. Res.* 50, 2 (2012), 495–540.
[59] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. 1188–1196.
[60] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.
[61] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'16)*. 3818–3824.
[62] Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. 2015. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 1751–1754.
[63] Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 708–717.

[64] Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Detect rumor and stance jointly by neural multi-task learning. In *Companion of the The Web Conference 2018 on The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 585–593.

[65] Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on Twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 1980–1989.

[66] Panagiotis Metaxas and Samantha Finn. 2017. The infamous "pizzagate" conspiracy theory: Insights from a TwitterTrails investigation. Retrieved on 2019 from http://cs.wellesley.edu~pmetaxas/Research/The_infamous_Pizzagate_conspiracy_theory_Insight_from_a_TwitterTrails_investigation.pdf.

[67] Rada Mihalcea and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Association for Computational Linguistics, 309–312.

[68] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[69] Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. 2012. Tweeting is believing?: Understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. ACM, 441–450.

[70] Alessandro Moshitti. 2006. Efficient convolution kernels for dependency and constituent syntactic trees. In *Proceedings of the European Conference on Machine Learning (ECML'06)*, Vol. 4212. Springer, 318–329.

[71] Subhabrata Mukherjee and Gerhard Weikum. 2015. Leveraging joint interactions for credibility analysis in news communities. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 353–362.

[72] Galit Nahari, Aldert Vrij, and Ronald P. Fisher. 2012. Does the truth come out in the writing? Scan as a lie detection tool. *Law Hum. Behav.* 36, 1 (2012), 68.

[73] Mark E. J. Newman. 2002. Spread of epidemic disease on networks. *Physical Review E* 66, 1 (2002), 016128.

[74] Matthew L. Newman, James W. Pennebaker, Diane S. Berry, and Jane M. Richards. 2003. Lying words: Predicting deception from linguistic styles. *Pers. Soc. Psychol. Bull.* 29, 5 (2003), 665–675.

[75] Nam P. Nguyen, Guanhua Yan, My T. Thai, and Stephan Eidenbenz. 2012. Containment of misinformation spread in online social networks. In *Proceedings of the 4th Annual ACM Web Science Conference*. ACM, 213–222.

[76] Myle Ott, Claire Cardie, and Jeffrey T. Hancock. 2013. Negative deceptive opinion spam. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*. 497–501.

[77] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. 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-Volume 1*. Association for Computational Linguistics, 309–319.

[78] Emory Paine. 2015. The next step: Social media and the evolution of journalism. Honors Theses 53. College of Arts and Sciences at Salem State University.

[79] Yiangos Papanastasiou. 2017. Fake news propagation and detection: A sequential model. *SSRN:3028354* (2017). https://doi.org/10.2139/ssrn.3028354

[80] James W. Pennebaker. 2001. Linguistic inquiry and word count: LIWC 2001 (2001), 71.

[81] Gordon Pennycook and David G. Rand. 2018. Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *SSRN:3023545* (2018). https://doi.org/10.2139/ssrn.3023545

[82] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 3391–3401. https://www.aclweb.org/anthology/C18-1287.

[83] Andrew Perrin. 2015. Social media usage. *Pew Research Center* (2015), 52–68. Retrieved on 2019 from https://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015.

[84] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2018. CredEye: A credibility lens for analyzing and explaining misinformation. In *Companion Proceedings of the The Web Conference (WWW'18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 155–158. https://doi.org/10.1145/3184558.3186967

[85] Kashyap Popat, Subhabrata Mukherjee, Jannik StrÃűtgen, and Gerhard Weikum. 2017. Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media. In *Proceedings of the 26th International Conference on World Wide Web Companion (WWW'17 Companion)*. ACM Press, New York, NY, 1003–1012. DOI: https://doi.org/10.1145/3041021.3055133

[86] Stephen Porter and John C. Yuille. 1996. The language of deceit: An investigation of the verbal clues to deception in the interrogation context. *Law Hum. Behav.* 20, 4 (1996), 443.

[87] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 231–240. https://www.aclweb.org/anthology/P18-1022.

[88] Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1589–1599.

[89] Feng Qian, Chengyue Gong, Karishma Sharma, and Yan Liu. 2018. Neural user response generator: Fake news detection with collective user intelligence. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18)*. International Joint Conferences on Artificial Intelligence Organization, 3834–3840. DOI: https://doi.org/10.24963/ijcai.2018/533

[90] Walter Quattrociocchi, Antonio Scala, and Cass R. Sunstein. 2016. Echo chambers on Facebook. *SSRN: 2795110* (2016). https://doi.org/10.2139/ssrn.2795110

[91] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2921–2927.

[92] Paul Rayson, Andrew Wilson, and Geoffrey Leech. 2001. Grammatical word class variation within the British National Corpus sampler. *Lang. Comput.* 36, 1 (2001), 295–306.

[93] Arne Roets et al. 2017. "Fake news": Incorrect, but hard to correct. The role of cognitive ability on the impact of false information on social impressions. *Intelligence* 65 (2017), 107–110.

[94] Jon Roozenbeek and Sander van der Linden. 2018. The fake news game: Actively inoculating against the risk of misinformation. *J, Risk Res,* (2018), 1–11. https://doi.org/10.1080/13669877.2018.1443491

[95] Victoria L. Rubin, Niall J. Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? Using satirical cues to detect potentially misleading news. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'16)*. 7–17.

[96] Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. CSI: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 797–806.

[97] David E. Rumelhart, Geoffrey E. Hinton, Ronald J. Williams, et al. 1988. Learning representations by back-propagating errors. *Cogn. Model.* 5, 3 (1988), 1.

[98] Alireza Saberi, Mojtaba Vahidi, and Behrouz Minaei Bidgoli. 2007. Learn to detect phishing scams using learning and ensemble? methods. In *Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology Workshops*. IEEE Computer Society, 311–314.

[99] Avinoam Sapir. 1987. The LSI course on scientific content analysis (SCAN). *Laboratory for Scientific Interrogation, Phoenix, AZ*.

[100] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. FakeNewsNet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286* (2018).

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

[102] Craig Silverman. 2015. Lies, damn lies and viral content. Tow Center for Digital Journalism Reports, Columbia University. Columbia University, NYC, New York. https://doi.org/10.7916/D8Q81RHH

[103] Craig Silverman, Lauren Strapagiel, Hamza Shaban, Ellie Hall, and Jeremy Singer-Vine. 2016. Hyperpartisan Facebook pages are publishing false and misleading information at an alarming rate. Retrieved on 2019 from https://www.buzzfeednews.com/article/craigsilverman/partisan-fb-pages-analysis.

[104] Aaron Smith and Monica Anderson. 2018. Social media use in 2018. Retrieved 2019 from https://www.pewinternet.org/2018/03/01/social-media-use-in-2018/.

[105] Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods In Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 1201–1211.

[106] Richard Socher, Cliff C. Lin, Chris Manning, and Andrew Y. Ng. 2011. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML'11)*. 129–136.

[107] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 1631–1642.

[108] V. S. Subrahmanian, Amos Azaria, Skylar Durst, Vadim Kagan, Aram Galstyan, Kristina Lerman, Linhong Zhu, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. The DARPA Twitter bot challenge. *Computer* 49, 6 (2016), 38–46.

[109] Eugenio Tacchini, Gabriele Ballarin, Marco L. Della Vedova, Stefano Moret, and Luca de Alfaro. 2017. Some like it hoax: Automated fake news detection in social networks. In *Proceedings of the Second Workshop on Data Science for Social Good. CEUR Workshop Proceedings*, Vol. 1960.

[110] Misako Takayasu, Kazuya Sato, Yukie Sano, Kenta Yamada, Wataru Miura, and Hideki Takayasu. 2015. Rumor diffusion and convergence during the 3.11 earthquake: A Twitter case study. *PLoS ONE* 10, 4 (2015), e0121443.

[111] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: A large-scale dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1 (Long Papers). Association for Computational Linguistics, New Orleans, 809–819. https://doi.org/10.18653/v1/N18-1074

[112] Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. 2018. Fake news detection in social networks via crowd signals. In *Companion of the The Web Conference 2018 on The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 517–524.

[113] Udo Undeutsch. 1984. Courtroom evaluation of eyewitness testimony. *Appl. Psychol.* 33, 1 (1984), 51–66.

[114] Prashanth Vijayaraghavan, Soroush Vosoughi, and Deb Roy. 2017. Twitter demographic classification using deep multi-modal multi-task learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 478–483.

[115] Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. 18–22.

[116] Svitlana Volkova and Jin Yea Jang. 2018. Misleading or falsification: Inferring deceptive strategies and types in online news and social media. In *Companion of the The Web Conference 2018 on The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 575–583.

[117] Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 647–653.

[118] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 422–426. DOI: https://doi.org/10.18653/v1/P17-2067

[119] Claire Wardle. 2017. Fake news. It's complicated. Retrieved on 2019 from https://firstdraftnews.org/fake-news-complicated/.

[120] Ke Wu, Song Yang, and Kenny Q. Zhu. 2015. False rumors detection on sina weibo by propagation structures. In *Proceedings of the 2015 IEEE 31st International Conference on Data Engineering (ICDE'15)*. IEEE, 651–662.

[121] Liang Wu and Huan Liu. 2018. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining (WSDM'18)*. ACM, New York, NY, 637–645. DOI: https://doi.org/10.1145/3159652.3159677

[122] Fan Yang, Yang Liu, Xiaohui Yu, and Min Yang. 2012. Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*. ACM, 13.

[123] Wen-tau Yih, Xiaodong He, and Christopher Meek. 2014. Semantic parsing for single-relation question answering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 643–648.

[124] YouGov. 2017. C4 study reveals only 4% surveyed can identify true or fake news. Retrieved from http://www.channel4.com/info/press/news/c4-study-reveals-only-4- surveyed-can-identify-true-or-fake-news.

[125] Sixie Yu, Yevgeniy Vorobeychik, and Scott Alfeld. 2018. Adversarial classification on social networks. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 211–219.

[126] Savvas Zannettou, Michael Sirivianos, Jeremy Blackburn, and Nicolas Kourtellis. 2018. The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *arXiv preprint arXiv:1804.03461* (2018).

[127] Min Zhang, GuoDong Zhou, and Aiti Aw. 2008. Exploring syntactic structured features over parse trees for relation extraction using kernel methods. *Inf. Process. Manage.* 44, 2 (2008), 687–701.

[128] Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1395–1405.

[129] Ke Zhou, Hongyuan Zha, and Le Song. 2013. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *Artificial Intelligence and Statistics*. 641–649.

[130] Lina Zhou, Judee K. Burgoon, Jay F. Nunamaker, and Doug Twitchell. 2004. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group Dec. Negot.* 13, 1 (2004), 81–106.

[131] Melissa Zimdars. 2016. False, Misleading, Clickbait-Y, and Satirical 'News' Sources. Retrieved 2019 from https://21stcenturywire.com/wpcontent/uploads/2017/02/2017-DR-ZIMDARS-False-Misleading-Clickbait-y-and-Satirical-%E2%80%9CNews%E2%80%9D-Sources-Google-Docs.pdf.

[132] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Comput. Surv.* 51, 2 (2018), 32.

[133] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS ONE* 11, 3 (Mar. 2016), e0150989. DOI: https://doi.org/10.1371/journal.pone.0150989

Received July 2018; revised December 2018; accepted December 2018