A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities

XINYI ZHOU and REZA ZAFARANI, Syracuse University

The explosive growth in fake news and its erosion to democracy, justice, and public trust has increased the demand for fake news detection and intervention. This survey reviews and evaluates methods that can detect fake news from four perspectives: the false *knowledge* it carries, its writing *style*, its *propagation* patterns, and the credibility of its *source*. The survey also highlights some potential research tasks based on the review. In particular, we identify and detail related fundamental theories across various disciplines to encourage interdisciplinary research on fake news. It is our hope that this survey can facilitate collaborative efforts among experts in computer and information sciences, social sciences, political science, and journalism to research fake news, where such efforts can lead to fake news detection that is not only efficient but, more importantly, explainable.

CCS Concepts: • **Human-centered computing** → *Collaborative and social computing theory, concepts and paradigms*; *Empirical studies in collaborative and social computing*; • **Computing methodologies** → *Natural language processing*; *Machine learning*; • **Security and privacy** → *Social aspects of security and privacy*; • **Applied computing** → *Sociology*; *Computer forensics*;

Additional Key Words and Phrases: Fake news, news verification, disinformation, misinformation, fact-checking, knowledge graph, deception detection, information credibility, social media

ACM Reference format:
Xinyi Zhou and Reza Zafarani. 2020. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Comput. Surv.* 53, 5, Article 109 (September 2020), 40 pages.
https://doi.org/10.1145/3395046

1 INTRODUCTION

Fake news is now viewed as one of the greatest threats to democracy, journalism, and freedom of expression. It has weakened public trust in governments, and its potential impact on the contentious "Brexit" referendum and the equally divisive 2016 U.S. presidential election—which it might have affected [Pogue 2017]—is yet to be realized [Allcott and Gentzkow 2017; Zafarani et al. 2019; Zhou et al. 2019b]. The reach of fake news was best highlighted during the critical months of the 2016 U.S. presidential election campaign. During that period, the top 20 frequently discussed fake election stories generated 8,711,000 shares, reactions, and comments on Facebook—ironically, more than the 7,367,000 for the top 20 most-discussed election stories posted by 19 major news websites [Silverman 2016]. Research has shown that compared to the truth, fake news on Twitter

Authors' addresses: X. Zhou and R. Zafarani, Data Lab, EECS Department, Syracuse University, Syracuse, NY 13244; emails: {zhouxinyi, reza}@data.syr.edu.
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2020 Association for Computing Machinery.
0360-0300/2020/09-ART109 $15.00
https://doi.org/10.1145/3395046

المنارة للاستشارات
www

is typically retweeted by many more users and spreads far more rapidly, especially for political news [Vosoughi et al. 2018]. Our economies are not immune to the spread of fake news either, with fake news being connected to stock market fluctuations and large trades. For example, fake news claiming that Barack Obama, the 44th President of the United States, was injured in an explosion wiped out $130 billion in stock value [Rapoza 2017]. These events and losses have motivated fake news research and sparked the discussion around fake news, as observed by skyrocketing usage of terms such as *post-truth*—selected as the international word of the year by Oxford Dictionaries in 2016 [Wang 2016].

Although fake news is not a new phenomenon [Tandoc et al. 2018], questions such as why it has emerged as a global topic of interest and why it is attracting increasingly more public attention are particularly relevant at this time. The leading cause is that fake news can be created and published online faster and cheaper when compared to traditional news media such as newspapers and television [Shu et al. 2017]. The rise of social media and its popularity also plays an essential role in this surge of interest [Olteanu et al. 2019; Zafarani et al. 2014]. As of August 2018, around two-thirds (68%) of Americans get their news from social media.[1] With the existence of an *echo chamber effect* on social media, biased information is often amplified and reinforced [Jamieson and Cappella 2008]. As an ideal platform to accelerate fake news dissemination, social media breaks the physical distance barrier among individuals; provides rich platforms to share, forward, vote, and review; and encourages users to participate and discuss online news. This surge of activity around online news can lead to grave repercussions and substantial potential political and economic benefits. Such generous benefits encourage malicious entities to create, publish, and spread fake news.

Take the dozens of "well-known" teenagers in the Macedonian town of Veles as an example of users who created fake news for millions on social media and became wealthy by penny-per-click advertising during the U.S. presidential election. As reported by NBC, each individual "has earned at least $60,000 in the past six months—far outstripping their parents' income and transforming his prospects in a town where the average annual wage is $4,800" [Smith and Banic 2016]. The tendency of individuals to overestimate the benefits associated with disseminating fake news rather than its costs, as the *valence effect* indicates [Jones and McGillis 1976], further attracts individuals to engage in fake news activities. Clearly, when governments, parties, and business tycoons are standing behind fake news generation, seeking its tempting power and profits, there is a greater motivation and capability to make fake news more persuasive and indistinguishable from truth to the public. But, how can fake news gain public trust?

Social and psychological factors play an important role in fake news gaining public trust and further facilitate the spread of fake news. For instance, humans have been proven to be irrational and vulnerable when differentiating between truth and falsehood while overloaded with deceptive information. Studies in social psychology and communications have demonstrated that human ability to detect deception is only slightly better than chance: typical accuracy rates are in the 55% to 58% range, with a mean accuracy of 54% over 1,000 participants in more than 100 experiments [Rubin 2010]. The situation is more critical for fake news than other types of information. For news, where one expects authenticity and objectivity, it is relatively easier to gain public trust. In addition, individuals tend to trust fake news after repeated exposures (*validity effect* [Boehm 1994]), or if it confirms their preexisting beliefs (*confirmation bias* [Nickerson 1998]) or attitudes (*selective exposure* [Freedman and Sears 1965; Metzger et al. 2015]), or if it pleases them (*desirability bias* [Fisher 1993]). *Peer pressure* can also at times "control" our perception and behavior (e.g., *bandwagon effect* [Leibenstein 1950]).

[1] https://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/.

المنارة للاستشارات
www

Table 1. Comparison Between Concepts Related to Fake News

Concept	Authenticity	Intention	News?
Deceptive news	Non-factual	Mislead	Yes
False news	Non-factual	Undefined	Yes
Satire news	Non-unified[2]	Entertain	Yes
Disinformation	Non-factual	Mislead	Undefined
Misinformation	Non-factual	Undefined	Undefined
Cherry-picking	Commonly factual	Mislead	Undefined
Clickbait	Undefined	Mislead	Undefined
Rumor	Undefined	Undefined	Undefined

The many perspectives on what fake news is, what characteristics and nature fake news or those who disseminate it share, and how fake news can be detected motivate the need for a comprehensive introduction and in-depth analysis, which this survey aims to develop. In addition, this survey aims to attract researchers within general areas of data mining, machine learning (ML), graph mining, natural language processing (NLP), and information retrieval (IR). More importantly, it is our hope to boost collaborative efforts among experts in computer and information sciences, political science, journalism, social sciences, psychology, and economics to study fake news, where such efforts can lead to fake news detection that is not only efficient but, more importantly, explainable [Zafarani et al. 2019; Zhou et al. 2019b].

To achieve these goals, we first discuss the ways to define fake news (Section 1.1) and summarize related fundamental theories across disciplines (e.g., in social sciences and economics) that can help study fake news (Section 1.2). Before further specification, we present an overview of this survey in Section 1.3.

1.1 What Is Fake News?

There has been no universal definition for fake news, even in journalism. A clear and accurate definition helps lay a solid foundation for fake news analysis and evaluating related studies. Here we (I) distinguish between several concepts that frequently co-occur or have overlaps with fake news; (II) present a broad and a narrow definition for the term *fake news*, justifying each definition; and (III) further discuss the potential research problems raised by such definitions.

I. Related concepts. Existing studies often connect fake news to terms and concepts such as *deceptive news* [Allcott and Gentzkow 2017; Lazer et al. 2018; Shu et al. 2017], *false news* [Vosoughi et al. 2018], *satire news* [Rubin et al. 2015; Tandoc et al. 2018; Wardle 2017], *disinformation* [Kshetri and Voas 2017; Wardle 2017], *misinformation* [Kucharski 2016; Wardle 2017], *cherry-picking* [Asudeh et al. 2020], *clickbait* [Chen et al. 2015], and *rumor* [Zubiaga et al. 2018]. Based on how these terms and concepts are defined, we can distinguish one from the others based on three characteristics: (i) *authenticity* (containing any non-factual statement or not), (ii) *intention* (aiming to mislead or entertain the public), and (iii) *whether the information is news*. Table 1 summarizes these related concepts based on these characteristics. For example, disinformation is false information (news or not-news) with a malicious intention to mislead the public.

II. Defining fake news. Challenges of fake news research start from defining fake news. To date, no universal definition is provided for fake news, where it has been looked upon as "a news article

[2]For example, Golbeck et al. [2018] regard satire news as "factually incorrect," whereas Tandoc et al. [2018] state that "where parodies differ from satires is their use of non-factual information to inject humor."

المنارة للاستشارات
www

that is intentionally and verifiably false" [Allcott and Gentzkow 2017; Shu et al. 2017] (deceptive news), or "a news article or message published and propagated through media, carrying false information regardless of the means and motives behind it," which overlaps with false news, disinformation [Kshetri and Voas 2017], misinformation [Kucharski 2016], satire news [Rubin et al. 2015], or even the stories that a person does not like (considered improper) [Golbeck et al. 2018]. Furthermore, what news is has become harder to define, as it can range from an account of a recent, interesting, and significant event to a dramatic account of something novel or deviant; in particular, "the digitization of news has challenged traditional definitions of news. Online platforms provide space for non-journalists to reach a mass audience" [Tandoc et al. 2018]. Under these circumstances, we first broadly define fake news as follows.

Definition 1 (Broad definition of fake news). Fake news is false news.

The definition of *news* broadly includes articles, claims, statements, speeches, and posts, among other types of information, related to public figures and organizations. It can be created by journalists and non-journalists. Such definition of news raises some social concerns (e.g., the term *fake news* should be "about more than news" and "about the entire information ecosystem" [Wardle 2017]). The broad definition aims to impose minimum constraints in accord with the current resources: it emphasizes information authenticity, purposefully adopts a broad definition for the term *news* [Vosoughi et al. 2018], and weakens the requirement for information intention due to the difficulty in obtaining the ground truth (true intention). This definition supports most existing fake-news-related studies and datasets, as provided by the existing fact-checking websites (Section 2.1 provides a detailed introduction). Current fake news datasets often provide ground truth for the authenticity of claims, statements, speeches, or posts related to public figures and organizations, yet limited information is provided on intentions.

We provide a more narrow definition of fake news, which satisfies the overall requirements for fake news as follows.

Definition 2 (Narrow Definition of Fake News). Fake news is intentionally false news published by a news outlet.

This narrow definition supports recent advancements in fake news studies [Allcott and Gentzkow 2017; Shu et al. 2017]. It addresses the public's perception of fake news, especially following the 2016 U.S. presidential election. Note that deceptive news is more harmful and less distinguishable than incautiously false news, as the former pretends to be truth to mislead the public better. The narrow definition emphasizes both news authenticity and intentions; it also ensures the posted information is news by investigating if its publisher is a news outlet (e.g., CNN and the New York Times). Often news outlets publish news in the form of articles with fixed components: a title, author(s), body text, image(s), and/or video(s) that include the claims made by, or about, public figures and organizations.

Both definitions require the authenticity of fake news to be false (i.e., being non-factual). As the goal is to provide a scientific definition for fake news, news falsity should be derived by comparing with *objective* facts and not with individual viewpoints (preferences). Hence, it is improper to consider fake news as articles that do not agree with individuals' or groups' interests or viewpoints, which is sometimes how the term *fake news* is used by the general public or in politics [Golbeck et al. 2018]. Such falsity can be assigned to the whole or part of the news content, or even to true news when subsequent events have rendered the original truth outdated (e.g., "Britain has control over fifty-six colonial countries"). In this general framework, a more comprehensive strategy for automatic fake news detection is needed, as the aforementioned fake news types emphasize various aspects of detection (see Section 6 for a discussion).

المنارة للاستشارات
www

III. Discussion. We have differentiated between fake and fake news-related terms based on three properties (authenticity, intention, and if it is news). We have also defined fake news as (i) deceptive news narrowly and as (ii) false news broadly. Questions are thus left, such as how can the authenticity and intention of the given information (news) be (manually or automatically) identified? Many domain experts and platforms have investigated ways to analyze news authenticity manually; however, how one can automatically assess news authenticity in an effective and explainable manner is still an open issue. We will detail both manual and automatic assessment of news authenticity (also known as fact-checking) in Section 2. To assess information (news) intention, analyzing the news writing style and its propagation patterns can be useful. First, the information (news) created to mislead or deceive the public intentionally (e.g., deceptive news) should look or sound "more persuasive" compared to the news without such intentions (e.g., satire news). Second, malicious users should play a part in the propagation of deceptive information (news) to enhance its social influence. Both news writing styles and propagation characteristics will be discussed with current methods in Sections 3 through 5. For intention analysis, often some level of manual news annotation is necessary. The accuracy of such annotations dramatically impacts automatic intention analysis within an ML framework. When the intentions behind non-factual information are determined, intervention strategies can be more appropriate and effective. For example, punitive measures should be taken against non-factual information and those who intentionally create it.

1.2 Fundamental Theories

Fundamental human cognition and behavior theories developed across various disciplines, such as social sciences and economics, provide invaluable insights for fake news analysis. These theories can introduce new opportunities for qualitative and quantitative studies of *big* fake news data [Zhou et al. 2019a]. These theories can also facilitate building well-justified and explainable models for fake news detection and intervention, which, to date, have been rarely available [Miller et al. 2017]. We have conducted a comprehensive literature survey across various disciplines and have identified well-known theories that can be potentially used to study fake news. These theories are provided in Table 2 along with short descriptions, which are related to either (I) the news itself or (II) its spreaders.

I. News-related theories. News-related theories reveal the possible characteristics of fake news content compared to true news content. For instance, theories have implied that fake news potentially differs from the truth in terms of, for example, writing style and quality (by *Undeutsch hypothesis*) [Undeutsch 1967], quantity such as word counts (by *information manipulation theory*) [McCornack et al. 2014], and sentiments expressed (by *four-factor theory*) [Zuckerman et al. 1981]. It should be noted that these theories, developed by forensic psychology, target deceptive statements or testimonies (i.e., disinformation) but not fake news, although these are similar concepts (see Section 1.1 for details). Thus, one research opportunity is to verify whether these attributes (e.g., information sentiment polarity) are statistically distinguishable among disinformation, fake news, and the truth, in particular, using big fake news data. However, these (discriminative) attributes identified can be used to automatically detect fake news using its writing style, where a typical study using supervised learning can be seen in the work of Zhou et al. [2019a]; we will provide further details in Section 3.

II. User-related theories. User-related theories investigate the characteristics of users involved in fake news activities, such as posting, forwarding, liking, and commenting. Fake news, unlike information such as fake reviews [Jindal and Liu 2008], can "attract" both malicious and normal users [Shao et al. 2018]. Malicious users (e.g., some social bots [Ferrara et al. 2016]) spread fake

المنارة للاستشارات
www

Table 2. Fundamental Theories in Social Sciences (Including Psychology and Philosophy) and Economics

		Theory	Phenomenon
News-Related Theories		Undeutsch hypothesis [Undeutsch 1967]	A statement based on a factual experience differs in content style and quality from that offantasy.
		Reality monitoring [Johnson and Raye 1981]	Actual events are characterized by higher levels of sensory-perceptual information.
		Four-factor theory [Zuckerman et al. 1981]	Lies are expressed differently in terms of arousal, behavior control, emotion, and thinking from truth.
		Information manipulation theory [McCornack et al. 2014]	Extreme information quantity often exists in deception.
User-Related Theories (User's Engagements and Roles in Fake News Activities)	**Social Impacts**	Conservatism bias [Basu 1997]	The tendency to revise one's belief insufficiently when presented with new evidence.
		Semmelweis reflex [Bálint and Bálint 2009]	Individuals tend to reject new evidence because it contradicts with established norms and beliefs.
		Echo chamber effect [Jamieson and Cappella 2008]	Beliefs are amplified or reinforced by communication and repetition within a closed system.
		Attentional bias [MacLeod et al. 1986]	An individual's perception is affected by his or her recurring thoughts atthe time.
		Validity effect [Boehm 1994]	Individuals tend to believe information is correct after repeated exposures.
		Bandwagon effect [Leibenstein 1950]	Individuals do something primarily because others are doing it.
		Normative influence theory [Deutsch and Gerard 1955]	The influence of others leading us to conform to be liked and accepted by them.
		Social identity theory [Ashforth and Mael 1989]	An individual's self-concept derives from perceived membership in a relevant social group.
		Availability cascade [Kuran and Sunstein 1999]	Individuals tend to adopt insights expressed by others when such insights are gaining more popularity within their social circles
	Self-Impact	Confirmation bias [Nickerson 1998]	Individuals tend to trust information that confirms their preexisting beliefs or hypotheses.
		Selective exposure [Freedman and Sears 1965]	Individuals prefer information that confirms their preexisting attitudes.
		Desirability bias [Fisher 1993]	Individuals are inclined to accept information that pleases them.
		Illusion of asymmetric insight [Pronin et al. 2001]	Individuals perceive their knowledge to surpass that of others.
		Naive realism [Ward et al. 1997]	The senses provide us with direct awareness of objects as they really are.
		Overconfidence effect [Dunning et al. 1990]	A person's subjective confidence in his judgments is reliably greater than the objective ones.
	Benefits	Prospecttheory [Kahneman and Tversky 2013]	People make decisions based on the value oflosses and gains rather than the outcome.
		Contrast effect [Hovland et al. 1957]	The enhancement or diminishment of cognition due to successive or simultaneous exposure to a stimulus oflesser or greater value in the same dimension.
		Valence effect [Frijda 1986]	People tend to overestimate the likelihood of good things happening rather than bad things.

المنارة للاستشارات
www

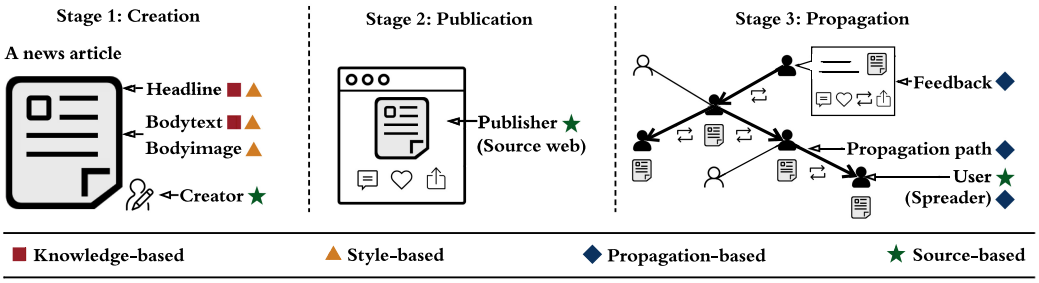


Fig. 1. Fake news life cycle and connections to the four fake news detection perspectives in this survey.

news often intentionally and are driven by *benefits* [Hovland et al. 1957; Kahneman and Tversky 2013]. Some normal users (which we denote as *vulnerable* normal users) can frequently and unintentionally spread fake news without recognizing the falsehood. Such vulnerability psychologically stems from *social impacts* and *self-impact*, where theories have been accordingly categorized and detailed in Table 2. Specifically, as indicated by the *bandwagon effect* [Leibenstein 1950], *normative influence theory* [Deutsch and Gerard 1955], *social identity theory* [Ashforth and Mael 1989], and *availability cascade* [Kuran and Sunstein 1999], to be liked and/or accepted by the community, normal users are encouraged to engage in fake news activities when many users have done so (i.e., *peer pressure*). One's trust to fake news and his or her unintentional spreading can be promoted as well when being *exposed* more to fake news (i.e., *validity effect*) [Boehm 1994], which often takes place due to the *echo chamber* effect on social media [Jamieson and Cappella 2008]. Such trust to fake news can be built when the fake news confirms one's *preexisting attitudes, beliefs or hypotheses* (i.e., *confirmation bias* [Nickerson 1998], *selective exposure* [Freedman and Sears 1965], and *desirability bias* [Fisher 1993]), which are often perceived to surpass that of others [Dunning et al. 1990; Pronin et al. 2001; Ward et al. 1997] and tend to be insufficiently revised when new refuting evidence is presented [Bálint and Bálint 2009; Basu 1997]. In such settings, strategies for intervening fake news from a user perspective (more discussions on fake news intervention are presented in Section 6) should be cautiously designed for users with different levels of credibility or intentions, even though they might all engage in the same fake news activity. For instance, it is reasonable to intervene with the spread of fake news by penalizing (e.g., removing) malicious users, but not for normal accounts. Instead, education and personal recommendations of true news articles and refuted fake ones can be helpful for normal users [Vo and Lee 2018]. Such recommendations should not only cater to the topics that the users want to read but should also capture topics that users are most gullible to. In Section 5, we will provide the path for utilizing these theories, such as quantifying social and self-impact, to enhance fake news research by identifying user intent and evaluating user credibility.

Meanwhile, we should point out that clearly understanding the potential roles that the fundamental theories listed in Table 2 can play in fake news research requires further in-depth investigations of interdisciplinary nature.

1.3 An Overview of This Survey

We have defined fake news (Section 1.1) and presented relevant fundamental theories in various disciplines (Section 1.2). The rest of this survey is organized as follows. We detail the detection of fake news from four perspectives (Figure 1 presents an overview): (I) *knowledge-based* methods

المنارة للاستشارات
www

(Section 2), which detect fake news by verifying if the knowledge within the news content (text) is consistent with facts (true knowledge); (II) *style-based* methods (Section 3), which are concerned with how fake news is written (e.g., if it is written with extreme emotions); (III) *propagation-based* methods (Section 4), where they detect fake news based on how it spreads online; and (IV) *source-based* methods (Section 5), which detect fake news by investigating the credibility of news sources at various stages (being created, published online, and spread on social media). In Section 6, we discuss open issues in current fake news studies and in fake news detection. We highlight six potential research tasks, hoping to facilitate the development of fake news research. We conclude in Section 7.

Comparison to related surveys. Our survey varies from related surveys from three perspectives. First, we discuss the ways fake news is defined in the current fake news research and its harmfulness to the public. We detail how fake news is related to terms such as *deceptive news*, *false news*, *satire news*, *disinformation*, *misinformation*, *cherry-picking*, *clickbait*, and *rumor*. Compared to related surveys and forums that often provide a specific definition for fake news, this survey highlights the challenges of defining fake news and introduces both a narrow and a broad definition for it.

Second, although recent studies have highlighted the importance of multidisciplinary fake news research [Lazer et al. 2018], we provide a path toward it by conducting an extensive literature survey across various disciplines, identifying a comprehensive list of well-known theories. We demonstrate how these theories relate to fake news and its spreaders and illustrate technical methods utilizing these theories both in fake news detection and intervention.

Third, for fake news detection, current surveys have mostly limited their scope to reviewing research from a certain perspective (or within a certain research area, e.g., NLP [Oshikawa et al. 2018] and data mining [Shu et al. 2017]). These surveys generally classify fake news detection models by the types of (deep) ML methods used [Oshikawa et al. 2018] or by whether they utilize social context information [Shu et al. 2017]. In our survey, we categorize automatic fake news detection methods from four perspectives: *knowledge* (Section 2), *style* (Section 3), *propagation* (Section 4), and *source* (Section 5). Reviewing and organizing fake news detection studies in such a way allows analyzing both *news content* (mainly Sections 2 and 3) and the *medium* (often social media) on which the news spreads (Sections 4 and 5), where fake news detection can be defined as a *probabilistic/regression problem* linked to, for example, *entity resolution* and *link prediction* tasks (Section 2), or a *classification problem* that relies on, for example, *feature engineering* and *(text and graph) embedding* techniques (Sections 3–5). In our survey of fake news detection, *patterns* of fake news in terms of its content (text and images, see Figures 7 and 8) or how it propagates (see Figures 10 and 15) are revealed, *algorithms* and *model architectures* are presented (e.g., Figures 6, 11, and 12–14), and *performance* of various fake news detection methods are compared (e.g., Table 6). We point out that our survey focuses more on how to construct a fake news dataset, such as, ground truth data, and the possible sources to obtain such ground truth (e.g., Section 2.1), rather than detailing existing datasets, which have been provided in past surveys [Oshikawa et al. 2018; Shu et al. 2017]. Nevertheless, we acknowledge the contributions to automatic fake news detection by these existing datasets (e.g., CREDBANK [Mitra and Gilbert 2015], LIAR [Wang 2017], FakeNewsNet [Shu et al. 2018], FakevsSatire [Golbeck et al. 2018], NELA-GT-2018 [Nørregaard et al. 2019], FEVER [Thorne et al. 2018], PHEME [Kochkina et al. 2018], and Emergent [Ferreira and Vlachos 2016]) and systems that can be used for building datasets (e.g., ClaimBuster [Hassan et al. 2017b], XFake [Yang et al. 2019], Hoaxy [Shao et al. 2016], MediaRank [Ye and Skiena 2019], Botometer [Davis et al. 2016], and RumorLens [Resnick et al. 2014]).

المنارة للاستشارات
www

Table 3. Comparison Among Expert-Based Fact-Checking Websites

Website	Topics Covered	Content Analyzed	Assessment Labels
PolitiFact[3]	American politics	Statements	True; Mostly true; Half true; Mostly false; False; Pants on fire
The Washington Post Fact Checker[4]	American politics	Statements and claims	One pinocchio; Two pinocchio; Three pinocchio; Four pinocchio; The Geppetto checkmark; An upside-down Pinocchio; Verdict pending
FactCheck[5]	American politics	TV ads, debates, speeches, interviews, and news	True; No evidence; False
Snopes[6]	Politics and other social and topical issues	News articles and videos	True; Mostly true; Mixture; Mostly false; False; Unproven; Outdated; Miscaptioned; Correct attribution; Misattributed; Scam; Legend
TruthOrFiction[7]	Politics, religion, nature, aviation, food, medical, etc.	Email rumors	Truth; Fiction; etc.
FullFact[8]	Economy, health, education, crime, immigration, law	Articles	Ambiguity (no clear labels)
HoaxSlayer[9]	Ambiguity	Articles and messages	Hoaxes, scams, malware, bogus warning, fake news, misleading, true, humor, spams, etc.
GossipCop[10]	Hollywood and celebrities	Articles	0–10 scale, where 0 indicates completely fake news and 10 indicates completely true news

2 KNOWLEDGE-BASED FAKE NEWS DETECTION

When detecting fake news from a knowledge-based perspective, one often uses a process known as *fact-checking*. Fact-checking, initially developed in journalism, aims to assess news authenticity by comparing the knowledge extracted from to-be-verified news content (e.g., its claims or statements) with known facts. In this section, we will discuss traditional fact-checking (also known as manual fact-checking) and how it can be incorporated into automatic means to detect fake news (i.e., automatic fact-checking).

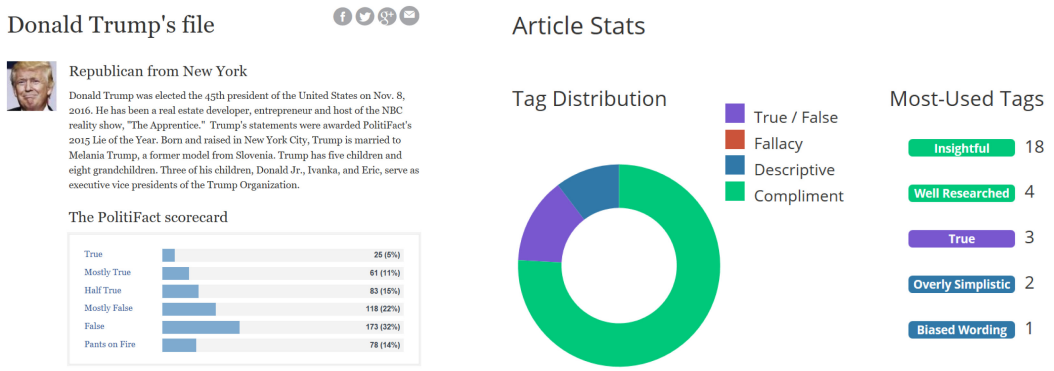
2.1 Manual Fact-Checking

Broadly speaking, manual fact-checking can be divided into (I) expert-based and (II) crowd-sourced fact-checking.

I. Expert-based manual fact-checking. Expert-based fact-checking relies on domain experts as *fact-checkers* to verify the given news contents. Expert-based fact-checking is often conducted by a small group of highly credible fact-checkers, is easy to manage, and leads to highly accurate results, but it is costly and poorly scales with the increase in the volume of the to-be-checked news contents.

[3] http://www.politifact.com/.
[4] https://www.factcheck.org/.
[5] https://www.washingtonpost.com/news/fact-checker.
[6] https://www.snopes.com/.
[7] https://www.truthorfiction.com/.
[8] https://fullfact.org/.
[9] http://hoax-slayer.com/.
[10] https://www.gossipcop.com.

المنارة للاستشارات
www



(a) (Expert-based) PolitiFact: the PolitiFact scorecard

(b) (Crowd-sourced) Fiskkit: the tag distribution

Fig. 2. Illustrations of manual fact-checking websites.

► *Expert-based fact-checking websites*. Recently, many websites have emerged to allow expert-based fact-checking better serve the public. We list and provide details on the well-known websites in Table 3. Some websites provide further information. For instance, *PolitiFact* provides "the PolitiFact scorecard," which presents statistics on the authenticity distribution of all statements related to a specific topic (see an example on Donald Trump, the 45th President of the United States, in Figure 2(a)). This information can provide the ground truth on the credibility of a topic [Zhang et al. 2018] and help identify check-worthy topics (see Section 6 for details) that require further scrutiny for verification. Another example is *HoaxSlayer*, which is different from most fact-checking websites that focus on information authenticity because it further classifies the articles and messages into categories such as hoaxes, spams, and fake news. Although the website does not provide clear definitions for these categories, its information can be potentially exploited as ground truth for comparative studies of fake news. In addition to the list provided here, a comprehensive list of fact-checking websites is provided by Reporters Lab at Duke University,[11] where more than 200 fact-checking websites across countries and languages are listed. Generally, these expert-based fact-checking websites can provide ground truth for the detection of fake news, particularly under the broad definition (Definition 1). Among these websites, *PolitiFact* and *GossipCop* have supported the development of fake news datasets that are publicly available (e.g., LIAR [Wang 2017] and FakeNewsNet [Shu et al. 2018]). The detailed expert-based analysis that these websites provide for checked contents (e.g., what is false and why is it false) carries invaluable insights for various aspects of fake news analysis, such as for *identifying check-worthy content* [Hassan et al. 2017a] and *explainable fake news detection* [Shu et al. 2019a] (see Section 6 for more discussion); however, to date, such insights have not been well utilized.

II. Crowd-sourced manual fact-checking. Crowd-sourced fact-checking relies on a large population of regular individuals acting as fact-checkers (i.e., the collective intelligence). Such large population of fact-checkers can be gathered within some common crowd-sourcing marketplaces such as Amazon Mechanical Turk, based on which CREDBANK [Mitra and Gilbert 2015], a publicly available large-scale fake news dataset, has been constructed. Compared to expert-based fact-checking, crowd-sourced fact-checking is relatively difficult to manage, less credible and accurate due to the political bias of fact-checkers and their conflicting annotations, and has better (although insufficient) scalability. Hence, in crowd-sourced fact-checking, one often needs to filter non-credible

[11] https://reporterslab.org/fact-checking/.

المنارة للاستشارات
www

users and resolve conflicting fact-checking results; both requirements become more critical as the number of fact-checkers grows. Nevertheless, crowd-sourcing platforms often allow fact-checkers to provide more detailed feedback (e.g., their sentiments or stances), which can be further explored in fake news studies.

► *Crowd-sourced fact-checking websites.* Unlike expert-based fact-checking, crowd-sourced fact-checking websites are still in early development. An example is *Fiskkit*,[12] where users can upload articles, provide ratings for sentences within articles, and choose tags that best describe the articles. The given sources of articles help distinguish the types of content (e.g., news vs. non-news) and determine its credibility (Section 5 provides the details). The tags categorized into multiple dimensions allow one to study the patterns across fake and non-fake news articles (see Figure 2(b) for an example). Although crowd-sourced fact-checking websites are not many, we believe that more crowd-sourced platforms or tools will arise as major web and social media websites start to realize their importance in identifying fake news (e.g., Google,[13] Facebook,[14] Twitter,[15] and Sina Weibo[16]).

2.2 Automatic Fact-Checking

Manual fact-checking does not scale with the volume of newly created information, especially on social media. To address scalability, automatic fact-checking techniques have been developed, heavily relying on IR, NLP, and ML techniques, as well as on network/graph theory [Cohen et al. 2011]. To review these techniques, the following definition presents a unified standard representation of knowledge is first presented that can be automatically processed by machines and has been widely adopted in related studies [Nickel et al. 2016].

Definition 3 (Knowledge). A set of `(Subject, Predicate, Object)` (SPO) triples extracted from the given information that well represent the given information.

For instance, the knowledge within the sentence "Donald Trump is the president of the U.S." can be (`DonaldTrump`, `Profession`, `President`). Based on the preceding representation of knowledge, we present the following widely accepted definitions for the key terms that often appear in automatic fact-checking literature for a better understanding [Ciampaglia et al. 2015; Dong et al. 2014; Shi and Weninger 2016].

Definition 4 (Fact). A fact is a knowledge (SPO triple) verified as truth.

Definition 5 (Knowledge base). A knowledge base (KB) is a set of facts.

Definition 6 (Knowledge Graph). A knowledge graph (KG) is a graph structure representing the SPO triples in a KB, where the entities (i.e., subjects or objects in SPO triples) are represented as nodes and relationships (i.e., predicates in SPO triples) are represented as edges.

Considering that a systematic approach for automatic fact-checking of news has to the best of our knowledge never been presented before, here we prioritize organizing the related research to clearly present the automatic news fact-checking process over presenting each related study in detail. The automatic fact-checking process is shown in Figure 3. It can be divided into two stages: fact extraction (a.k.a. *KB construction*, see Section 2.2.1) and fact-checking (a.k.a. *knowledge comparison*, see Section 2.2.2).

[12] http://fiskkit.com/.
[13] https://blog.google/topics/journalism-news/labeling-fact-check-articles-google-news/.
[14] https://newsroom.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/.
[15] https://blog.twitter.com/2010/trust-and-safety.
[16] http://service.account.weibo.com/ (sign in required).

المنارة للاستشارات
www

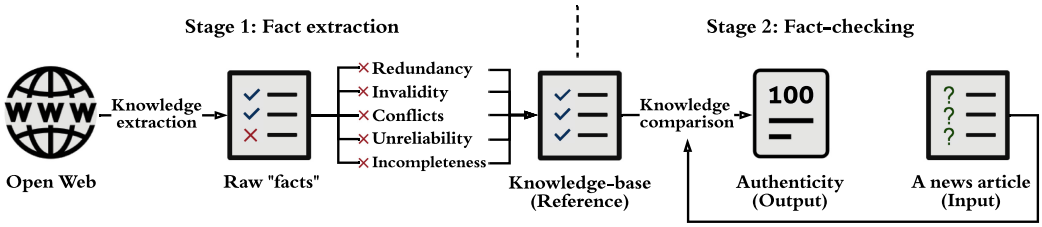


Fig. 3. Automatic news fact-checking process.

2.2.1 Fact Extraction. To collect facts and construct a KB (KG), knowledge is first extracted from the open web as raw "facts" that need further processing. Such process often refers to knowledge extraction or *relation extraction* [Nickel et al. 2016]. Knowledge extraction can be classified into *single-source* or *open-source* knowledge extraction. Single-source knowledge extraction, which relies on one comparatively reliable source (e.g., Wikipedia) to extract knowledge, is relatively efficient but often leads to incomplete knowledge (see related studies in the work of Auer et al. [2007], Bollacker et al. [2008], and Suchanek et al. [2007]). Open-source knowledge extraction aims to fuse knowledge from distinct sources, and hence, it is less efficient than single-source knowledge extraction but leads to more complete knowledge (see related studies in the work of Carlson et al. [2010], Dong et al. [2014], Nakashole et al. [2012], and Niu et al. [2012]). More recent studies in relation extraction can be seen in the work of Di et al. [2019], Lin et al. [2019], and Yu et al. [2019]. Finally, to form a KB (KG) from these extracted raw "facts," they need to be further cleaned up and completed by addressing the following issues:

- *Redundancy*: For example, `(DonaldJohnTrump, profession, President)` is redundant when having `(Donald- Trump, profession, President)` as `DonaldTrump` and `DonaldJohnTrump` match the same entity. The task to reduce redundancy is often referred to as *entity resolution* [Getoor and Machanavajjhala 2012], which is also known as *deduplication* or *record linkage* [Nickel et al. 2016] (e.g., see related studies such as Altowim et al. [2014], Bhattacharya and Getoor [2007], Christen [2008], Stoerts et al. [2016], and Whang and Garcia-Molina [2012]).
- *Invalidity*: The correctness of some facts depends on a specific time interval. For example, `(Britain, joinIn, EuropeanUnion)` has been outdated and should be updated. One way to address this issue is to allow facts to have beginning and ending dates [Bollacker et al. 2008], or one can *reify* current facts by adding extra assertions to them [Hoffart et al. 2013].
- *Conflicts*: For example, `(DonaldTrump, bornIn, NewYorkCity)` and `(DonaldTrump, bornIn, LosAngeles)` is a pair with conflicting knowledge. Conflicts can be resolved by multi-criteria decision-making (MCDM) methods [Deng 2015; Kang and Deng 2019; Pasi et al. 2019; Viviani and Pasi 2016].
- *Unreliability (low credibility)*: For example, the knowledge extracted from The Onion,[17] a satire news organization, is unreliable knowledge. Unreliable knowledge can be reduced by filtering low-credibility website(s) from which the knowledge is extracted. The credibility of website(s) can be obtained from resources such as NewsGuard[18] (expert-based) or systems such as MediaRank [Ye and Skiena 2019]. Section 5 provides more details.
- *Incompleteness*: Raw facts extracted from online resources, particularly using a single source, are far from complete. Hence, reliably inferring new facts based on existing facts, also

[17] https://www.theonion.com/.
[18] http://www.newsguardtech.com/.

المنارة للاستشارات
www

known as *KG completion*, is necessary to improve the KG being built. KG completion performs link prediction between entities, where the methods can be classified into three groups based on their assumptions: (i) *latent feature models*, which assume that the existence of KB triples is conditionally independent given latent features and parameters (e.g., RESCAL [Nickel et al. 2012], NTN [Socher et al. 2013], DistMult [Yang et al. 2014], TransE [Bordes et al. 2013], TransH [Wang et al. 2014], TransR [Lin et al. 2015], ComplEx [Trouillon et al. 2016], SimplE [Kazemi and Poole 2018], and ConMask [Shi and Weninger 2018]); (ii) *graph feature models*, which assume that the existence of triples is conditionally independent given observed graph features and parameters (e.g., the path ranking algorithm (PRA) [Lao and Cohen 2010]); and (iii) *Markov random field* (MRF) *models*, which assume that existing triples have local interactions [Nickel et al. 2016].

Note that instead of building a KB (KG) from scratch, one can rely on existing large-scale ones, such as YAGO [Hoffart et al. 2013; Suchanek et al. 2007], Freebase [Bollacker et al. 2008], NELL [Carlson et al. 2010], PATTY [Nakashole et al. 2012], DBpedia [Auer et al. 2007], Elementary/DeepDive [Niu et al. 2012], and Knowledge Vault [Dong et al. 2014].

2.2.2 Fact-Checking. To assess the authenticity of news articles, we need to compare the knowledge extracted from to-be-verified news content (i.e., SPO triples) with the facts (i.e., true knowledge). KBs (KGs) are suitable resources for providing ground truth for news fact-checking (i.e., we can reasonably assume that the existing triples in a KB (KG) represent facts). However, for non-existing triples, their authenticity relies on assumptions made—we list three common assumptions next—and we may need further inference:

- *Closed-world assumption*: Non-existing triples indicate false knowledge.
- *Open-world assumption*: Non-existing triples indicate unknown knowledge that can be either true or false.
- *Local closed-world assumption [Dong et al. 2014]*: The authenticity of non-existing triples can be determined by the following rule: let $T(s, p)$ denote the set of existing triples in the KB (KG) for subject s and predicate p. For any $(s, p, o) \notin T(s, p)$, if $|T(s, p)| > 0$, we say the triple is false; if $|T(s, p)| = 0$, its authenticity is unknown.

Generally, the fact-checking strategy for an SPO triple is to evaluate the *possibility* that the edge labeled `Predicate` exists from the node labeled `Subject` to the node representing `Object` in a KG. Specifically,

Step 1: *Entity locating.* `Subject` (Similarly, `Object`) is first matched with a node in the KG that represents the same entity as the `Subject` (similarly, `Object`), where *entity resolution* techniques (e.g., [Altowim et al. 2014; Bhattacharya and Getoor 2007; Trivedi et al. 2018]) can be used to identify proper matchings.

Step 2: *Relation verification.* The SPO triple is considered as truth if an edge labeled `Predicate` from the node representing `Subject` to the one representing `Object` exists in the KG. Otherwise, its authenticity is (i) false based on the aforementioned closed-world assumption or (ii) determined after *knowledge inference*.

Step 3: *Knowledge inference.* When the SPO triple does not exist in the KG, the probability for the edge labeled `Predicate` to exist from the node representing `Subject` to the one representing `Object` in the KG can be computed, for example, using *link prediction* methods such as semantic proximity [Ciampaglia et al. 2015], discriminative predicate path [Shi and Weninger 2016], or LinkNBed [Trivedi et al. 2018].

المنارة للاستشارات
www

We conclude this section by formally defining the problem of automatic news fact-checking in Problem 1 and discussing the potential research avenues in automatic news fact-checking in Section 2.2.3.

Problem 1 (Fact-checking). Assume that a to-be-verified news article is represented as a set of knowledge statements such as SPO triples $(s_i, p_i, o_i), i = 1, 2, \ldots, n$. *Let* G_{KB} *refer to a KG containing a set of facts (i.e., true knowledge) denoted by* $(s_{t_j}, p_{t_j}, o_{t_j}), j = 1, 2, \ldots, m$. *News fact-checking is to identify a function* $\mathcal{F}$ *that assigns an authenticity value* $A_i \in [0, 1]$ *to each corresponding* (s_i, p_i, o_i) *by comparing it with every* $(s_{t_j}, p_{t_j}, o_{t_j})$ *in the KG, where* $A_i = 1$ *(*$A_i = 0$*) indicates the triple is true (false). The final authenticity index* $A \in [0, 1]$ *of the to-be-verified news article is obtained by aggregating all* A_i*'s. To summarize,*

$$\begin{aligned} &\mathcal{F} : (s_i, p_i, o_i) \xrightarrow{G_{KB}} A_i, \\ &A = \mathcal{I}(A_1, A_2, \ldots, A_n), \end{aligned} \tag{1}$$

where $\mathcal{I}$ *is an aggregation function of choice (e.g., weighted or arithmetic average). The to-be-verified news article is true if* $A = 1$ *and is [completely] false if* $A = 0$. *Specifically, function* $\mathcal{F}$ *can be formulated as*

$$\mathcal{F}((s_i, p_i, o_i), G_{KB}) = P(\text{edge labeled } p_i \text{ linking } s_i' \text{ to } o_i' \text{ in } G_{KB}), \tag{2}$$

where $P(\cdot)$ *denotes the probability, and* s_i' *and* o_i' *are the matched entities to* s_i *and* o_i *in* G_{KB}*, respectively:*

$$s_i' = \arg\min_{s_{t_j}} |\mathcal{D}(s_i, s_{t_j})| < \theta, \quad o_i' = \arg\min_{o_{t_j}} |\mathcal{D}(o_i, o_{t_j})| < \theta. \tag{3}$$

In Equation (3), $\mathcal{D}(a, b)$ *measures the distance between entity* a *and entity* b. *Such distance can be computed by, for example, Jaccard distance directly or by cosine distance after entity embedding. When the distance between* a *and* b *is zero (i.e.,* $\mathcal{D}(a, b) = 0$*) or the distance is less than a certain threshold* θ *(i.e.,* $|\mathcal{D}(a, b)| < \theta$*), one can regard* a *as the same entity as* b.

2.2.3 Discussion. We have detailed fact extraction (i.e., KB/KG construction) and fact-checking (i.e., knowledge comparison), which are the two main components of automatic news fact-checking. There are some open issues and several potential research tasks. First, when collecting facts to construct KB (KG), one concern is the source(s) from which facts are extracted. In addition to traditional sources such as Wikipedia, some other sources, such as fact-checking websites that contain *expert analysis and justifications* for checked news content, might help provide high-quality domain knowledge. However, such sources have rarely been considered in current research. Second, we highlight the value of research on *dynamic KBs (KGs)* for news fact-checking that can automatically remove invalid knowledge and introduce new facts. Such properties are especially important due to news timeliness—news articles are often not about "common knowledge" but around recent events. Third, it has been verified that fake news spreads faster than true news [Vosoughi et al. 2018], which attaches great importance to fast news fact-checking to achieve *fake news early detection* (see Section 6 for a summary and discussion). Current research in building KBs (KGs) has focused on constructing KBs (KGs) with as many facts as possible. However, fast news fact-checking requires not only identifying parts of the to-be-verified news that is check-worthy (see Section 6 for a discussion on identifying check-worthy content) but also a KB (KG) that only stores as many "valuable" facts as possible (i.e., a *KB (KG) simplification process*).

3 STYLE-BASED FAKE NEWS DETECTION

Similar to knowledge-based fake news detection (Section 2), style-based fake news detection also focuses on analyzing the news content. However, knowledge-based methods mainly evaluate the authenticity of the given news, whereas style-based methods can assess news intention (i.e., is

المنارة للاستشارات
www

there an intention to mislead the public or not?). The intuition and assumption behind style-based methods is that malicious entities prefer to write fake news in a "special" style to encourage others to read and convince them to trust. Before discussing how such "special" content styles can be automatically identified, we first define fake news style in a way that facilitates use of ML.

Definition 7 (Fake News Style). A set of quantifiable characteristics (e.g., ML features) that can well represent fake news content and differentiate it from true news content.

Based on this definition for fake news style, style-based fake news detection is often formulated as a binary (or, at times, a multi-label) classification problem:

Problem 2 (Style-based fake news detection). Assume that a to-be-verified news article $\mathcal{N}$ can be represented as a set of k content features denoted by feature vector $\mathbf{f} \in \mathbb{R}^k$. The task to verify the news article based on its content style is to identify a function $\mathcal{S}$ such that

$$\mathcal{S} : \mathbf{f} \xrightarrow{TD} \hat{y}, \tag{4}$$

where $\hat{y} \in \{0(\texttt{true}), 1(\texttt{fake})\}$ is the predicted news label and $TD = \{(\mathbf{f}_l, y_l) \mid \mathbf{f}_l \in \mathbb{R}^k, y_l \in \{0, 1\}, l = 1 \ldots n\}$ is the training dataset. The training dataset helps estimate the parameters within $\mathcal{S}$ and consists of a set of n news articles represented by the same set of features ($\mathbf{f}_l$) with known news labels (y_l).

Hence, the performance of style-based fake news detection methods rely on (I) how well the style of news content (text and images) can be captured and represented (Section 3.1) and (II) how the classifier (model) is performing based on different news content representations (Section 3.2). In addition, we summarize (III) some verified patterns of fake news content style in Section 3.3 and provide (IV) our discussion on style-based methods in Section 3.4.

3.1 Style Representation

As provided in Definition 7, the content style is commonly represented by a set of quantifiable characteristics, often ML features. Generally, these features can be grouped into *textual features* (see Section 3.1.1) and *visual features* (see Section 3.1.2), representing news text and images, respectively.

3.1.1 News Text. Broadly speaking, textual features can be grouped into (I) *general features* and (II) *latent features*.

► *General textual features.* General textual features are often used to detect fake news within a traditional ML framework (detailed in Section 3.2.1). These features describe content style from (at least) four language levels: *lexicon*, *syntax*, *discourse*, and *semantic* [Conroy et al. 2015]. The main task at the lexicon level is to assess the frequency statistics of lexicons, which can be basically conducted using a bag-of-words (BOW) model [Zhou et al. 2019a]. At the syntax level, shallow syntax tasks are performed by part-of-speech (POS) taggers to assess POS (e.g., nouns and verbs) frequencies [Feng et al. 2012; Zhou et al. 2019a]. Deep syntax tasks are performed by probabilistic context-free grammar (PCFG) parse trees (Figure 4 presents an example) that enable assessing the frequencies of rewrite rules (i.e., productions) [Feng et al. 2012; Pérez-Rosas et al. 2017; Zhou et al. 2019a]. Four different encodings of rewrite rules can be considered for a PCFG parse tree [Feng et al. 2012]:

- r: Unlexicalized rewrite rules (i.e., all rewrite rules except for those with leaf nodes such as IN→"in");
- $r*$: Lexicalized rewrite rules (i.e., all rewrite rules);

المنارة للاستشارات
www

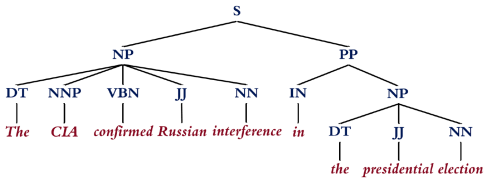


Fig. 4. PCFG parse tree for the sentence "The CIA confirmed Russian interference in the presidential election" in a fake news article (directly from Zhou et al. [2019a]). The lexicalized rewrite rules of this sentence are S→NP PP, NP→DT NNP VBN JJ NN, PP→IN NP, NP→DT JJ NN, DT→"the," NNP→"CIA," VBN→"confirmed," JJ→"Russian," NN→"interference," IN→"in," JJ→"presidential," and NN→"election."

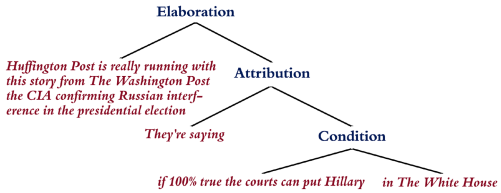


Fig. 5. Rhetorical structure for the partial content "Huffington Post is really running with this story from The Washington Post about the CIA confirming Russian interference in the presidential election. They're saying if 100% true, the courts can PUT HILLARY IN THE WHITE HOUSE!" in a fake news article (directly from Zhou et al. [2019a]). Here, one elaboration, one attribution, and one condition rhetorical relationship exist.

- $\hat{r}$: Unlexicalized rewrite rules with grandparent nodes (e.g., PPˆS→IN NP); and
- $\hat{r}*$: Lexicalized rewrite rules with grandparent nodes (e.g., INˆPP→"in").

At the discourse level, rhetorical structure theory (RST) and rhetorical parsing tools can be used to capture the frequencies of rhetorical relations among sentences as features [Karimi and Tang 2019; Zhou et al. 2019a] (Figure 5 presents an illustration). Finally, at a semantic level, such frequencies can be assigned to lexicons or phrases that fall into each psycho-linguistic category (e.g., those defined in linguistic inquiry and word count (LIWC) [Pérez-Rosas et al. 2017]), or that fall into each self-defined psycho-linguistic attribute. These attributes can be derived from experience, or be inspired by related deception theories (see news-related theories in Table 2 or in the work of Zhou et al. [2019a] as a typical interdisciplinary fake news study). Based on our investigation, such attributes and their corresponding computational features can be grouped along 10 dimensions: *quantity* [McCornack et al. 2014], *complexity*, *uncertainty*, *subjectivity* [Undeutsch 1967], *non-immediacy*, *sentiment* [Zuckerman et al. 1981], *diversity* [Undeutsch 1967], *informality* [Undeutsch 1967], *specificity* [Johnson and Raye 1981], and *readability* (Table 4), which are initially developed for identifying deception in computer-mediated communications [Fuller et al. 2009; Zhou et al. 2004b] and testimonies [Afroz et al. 2012], and recently used in fake news detection [Bond et al. 2017; Pérez-Rosas et al. 2017; Potthast et al. 2017; Zhou et al. 2019a].

It should be noted that "frequency" can be defined and computed in three ways. Assume that a corpus C contains p news articles $C = \{\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_p\}$ and a total of q words $W = \{w_1, w_2, \ldots, w_q\}$ (POS tags, rewrite rules, rhetorical relationships, etc.). x_j^i denotes the number of w_j appearing in $\mathcal{A}_i$. Then the "frequency" of w_j for news $\mathcal{A}_i$ can be

- *Absolute frequency* f_a, such as $f_a = x_j^i$;
- *Standardized frequency* f_s, which removes the impact of content length, such as $f_s = \frac{x_j^i}{\sum_j x_j^i}$ [Zhou et al. 2019a]; or
- *Relative frequency* f_r by using term frequency–inverse document frequency (TF-IDF), which further compares such frequency with that in other news articles in the corpus, such as $f_r = \frac{x_j^i}{\sum_j x_j^i} \ln p \frac{1}{\sum_i x_j^i}$ [Pérez-Rosas et al. 2017].

المنارة للاستشارات
www

Table 4. Semantic-Level Features in News Content

Attribute Type	Feature	[Zhou et al. 2004b]	[Fuller et al. 2009]	[Afroz et al. 2012]	[Siering et al. 2016]	[Zhang et al. 2016]	[Bond et al. 2017]	[Potthast et al. 2017]	[Perez-Rosas et al. 2017]	[Zhou et al. 2019a]
Quantity	# Characters			✓						✓
	# Words	✓	✓	✓	✓	✓				✓
	# Noun phrases	✓								
	# Sentences	✓	✓	✓	✓					✓
	# Paragraphs							✓		✓
Complexity	Average # characters per word	✓	✓	✓	✓					✓
	Average # words per sentence	✓	✓	✓	✓	✓				✓
	Average # clauses per sentence	✓			✓					
	Average # punctuations per sentence	✓	✓	✓	✓					
Uncertainty	#/% Modal verbs (e.g., "shall")	✓	✓	✓	✓					
	#/% Certainty terms (e.g., "never" and "always")	✓	✓	✓	✓		✓			✓
	#/% Generalizing terms (e.g., "generally" and "all")		✓							
	#/% Tentative terms (e.g., "probably")		✓	✓			✓			✓
	#/% Numbers and quantifiers			✓						
	#/% Question marks			✓						✓
Subjectivity	#/% Biased lexicons (e.g., "attack")									✓
	#/% Subjective verbs (e.g., "feel" and "believe")	✓				✓				
	#/% Report verbs (e.g., "announce")									✓
	#/% Factive verbs (e.g., "observe")									✓
Non-immediacy	#/% Passive voice	✓	✓							
	#/% Self reference: 1st person singular pronouns	✓	✓	✓	✓	✓				
	#/% Group reference: 1st person plural pronouns	✓	✓	✓	✓	✓				
	#/% Other reference: 2nd and 3rd person pronouns	✓	✓	✓	✓	✓				
	#/% Quotations			✓				✓		
Sentiment	#/% Positive words	✓	✓	✓	✓	✓	✓			✓
	#/% Negative words	✓	✓	✓	✓	✓	✓			✓
	#/% Anxiety/angry/sadness words						✓			✓
	#/% Exclamation marks			✓						✓
	Content sentiment polarity									✓
Diversity	Lexical diversity: #/% unique words or terms	✓	✓	✓	✓	✓				✓
	Content word diversity: #/% unique content words	✓	✓			✓				✓
	Redundancy: #/% unique function words	✓	✓	✓		✓				✓
	#/% Unique nouns/verbs/adjectives/adverbs									✓
Informality	#/% Typos (misspelled words)	✓			✓	✓				
	#/% Swear words/netspeak/assent/nonfluencies/fillers									✓
Specificity	Temporal/spatial ratio	✓	✓				✓			
	Sensory ratio	✓	✓		✓		✓			✓
	Causation terms		✓				✓			✓
	Exclusive terms		✓							
Readablity (e.g., Flesch-Kincaid and Gunning-Fog index)				✓				✓	✓	✓

The studies labeled with gray background color investigate news articles.

المنارة للاستشارات
www

Table 5. Visual Features in News Content (Defined by Jin et al. [2017])

Feature	Description
Visual Clarity Score	Distribution difference between the image set of a news article and that in the corpus
Visual Coherence Score	Average similarities between pairs of images in a news article
Visual Similarity Distribution Histogram	Distribution histogram of the similarity matrix of images in a news article
Visual Diversity Score	Weighted average of dissimilarities between image pairs in a news article
Visual Clustering Score	The number of image clusters in a news article

In general, TF-IDF can be applied at various language levels, and so can n-gram models, which enable capturing the sequence of words (POS tags, rewrite rules, etc.) [Feng et al. 2012; Pérez-Rosas et al. 2017].

► *Latent textual features.* Latent textual features are often used for news text embedding. Such an embedding can be conducted at the word level [Mikolov et al. 2013; Pennington et al. 2014], sentence level [Arora et al. 2016; Le and Mikolov 2014], or document level [Le and Mikolov 2014]; results are vectors representing a news article and can be directly used as the input to classifiers (e.g., SVMs) when predicting fake news within a traditional ML framework [Zhou et al. 2019a] (detailed in Section 3.2.1). Such embeddings (often at the word level) can be further incorporated into neural network architectures (e.g., convolutional neural networks (CNNs) [He et al. 2016; Huang et al. 2017; Krizhevsky et al. 2012; LeCun et al. 1989; Simonyan and Zisserman 2014; Szegedy et al. 2015], recurrent neural networks (RNNs) [Cho et al. 2014; Hochreiter and Schmidhuber 1997; Schuster and Paliwal 1997], and the Transformer [Devlin et al. 2018; Vaswani et al. 2017]) to predict fake news within a deep learning (DL) framework (see details in Section 3.2.2), where CNNs represent news text from a local to global view, and RNNs and the Transformer capture the sequences with news text. Theoretically, such latent representation can also be obtained by matrix or tensor factorization; current style-based fake news detection studies have rarely considered them, although a few studies focusing on the propagation aspects of fake news have considered these methods, which we will review later in Section 4.

3.1.2 News Images. Currently, not many studies exist on detecting fake news by exploring news images [Shu et al. 2017]. News images can be represented by handcrafted (i.e., non-latent) features, such as the visual features defined by Jin et al. [2017] (Table 5). However, to be further processed by neural networks such as VGG-16/19 [Simonyan and Zisserman 2014] to obtain a latent representation, each image is often embedded as a pixel matrix or tensor with size $width \times height \times \#channel(s)$, where channels can be the gray value (#channels=1) or RGB data (#channels=3).

3.2 Style Classification

We detail style-based fake news detection models that rely on traditional ML (in Section 3.2.1) and on DL [LeCun et al. 2015] (in Section 3.2.2), along with their performance results.

3.2.1 Traditional ML-Based Models. Within a traditional ML framework, news content is represented by using a set of manually selected (latent and non-latent) features, which can be extracted from news images, or text at various language levels (lexicon-, syntax-, semantic-, and discourse levels) [Feng et al. 2012; Pérez-Rosas et al. 2017; Zhou et al. 2019a]. ML models that can detect news type (e.g., true or fake) based on this representation can be supervised,

المنارة للاستشارات
www

Table 6. Feature Performance (Accuracy (Acc.) and $\mathbf{F_1}$-score) in Fake News Detection Using Traditional ML (RF and XGBoost Classifiers) [Zhou et al. 2019a]

Feature Group			PolitiFact data [Shu et al. 2018]				BuzzFeed data [Shu et al. 2018]			
			XGBoost		RF		XGBoost		RF	
			Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1
Non-Latent Features	Lexicon	BOWs (f_s)	0.856	0.858	0.837	0.836	0.823	0.823	0.815	0.815
		Unigram+bigram (f_r)	0.755	0.756	0.754	0.755	0.721	0.711	0.735	0.723
	Syntax	POS tags (f_s)	0.755	0.755	0.776	0.776	0.745	0.745	0.732	0.732
		Rewrite rules ($r*$, f_s)	**0.877**	**0.877**	0.836	0.836	0.778	0.778	0.845	0.845
		Rewrite rules ($r*$, f_r)	0.749	0.753	0.743	0.748	0.735	0.738	0.732	0.735
	Semantic	LIWC	0.645	0.649	0.645	0.647	0.655	0.655	0.663	0.659
		Theory-driven [Zhou et al. 2019a]	0.745	0.748	0.737	0.737	0.722	0.750	0.789	0.789
	Discourse	Rhetorical relationships	0.621	0.621	0.633	0.633	0.658	0.658	0.665	0.665
	Combination	[Zhou et al. 2019a]	0.865	0.865	**0.845**	**0.845**	**0.855**	**0.856**	**0.854**	**0.854**
Latent Features		Word2Vec [Mikolov et al. 2013]	0.688	0.671	0.663	0.667	0.703	0.714	0.722	0.718
		Doc2Vec [Le and Mikolov 2014]	0.698	0.684	0.712	0.698	0.615	0.610	0.620	0.615

Results show that (i) non-latent features can outperform latent ones, (ii) combining features across levels can outperform using single-level features, and (iii) the (standardized) frequencies of lexicons and rewrite rules better represent fake news content style and perform better (while being more time consuming to compute) than other feature groups.

semi-supervised, or unsupervised, where supervised methods (classifiers) have been mainly used for style-based fake news detection—for example, style-based methods have relied on SVMs [Feng et al. 2012; Pérez-Rosas et al. 2017], Random Forests (RF) [Zhou et al. 2019a], and XGBoost [Chen and Guestrin 2016; Zhou et al. 2019a].

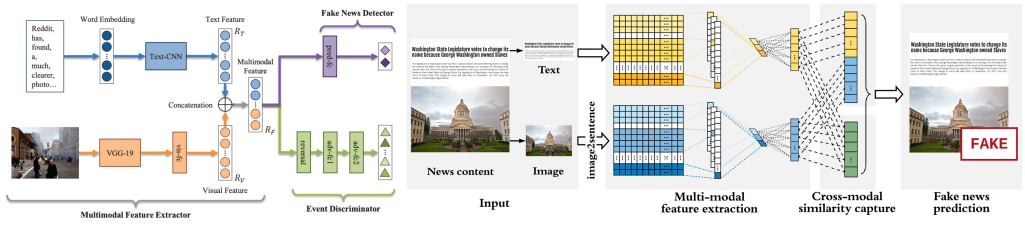
Within the same experimental setup as that in the work of Zhou et al. [2019a], the performance of (latent and non-latent) features at various levels is presented and compared in Table 6. Results indicate that when predicting fake news using a traditional ML framework, (i) non-latent features often outperform latent ones, (ii) combining features across levels can outperform using single-level features, and (iii) (standardized) frequencies of lexicons and rewrite rules better represent fake news content style and perform better (while being more time consuming to compute) than other feature groups.

It should be noted that as classifiers perform best for ML settings they were initially designed for, it is unjustified to determine algorithms that perform best for fake news detection (related discussions can be found in the work of Fernández-Delgado et al. [2014] and Kotsiantis et al. [2007]).

3.2.2 DL-Based Models. Within a DL framework, news content (text and/or images) is often first embedded at the word level [Mikolov et al. 2013] (for text), or as a pixel matrix or tensor (for images). Then, such an embedding is processed by a well-trained neural network (e.g., CNNs [He et al. 2016; Huang et al. 2017; Krizhevsky et al. 2012; LeCun et al. 1989; Szegedy et al. 2015] such as VGG-16/19 [Simonyan and Zisserman 2014] and Text-CNN [Kim 2014]; RNNs such as LSTMs [Hochreiter and Schmidhuber 1997], GRUs [Cho et al. 2014], and BRNNs [Schuster and Paliwal 1997]; and the Transformer [Devlin et al. 2018; Vaswani et al. 2017]) to extract latent textual and/or visual features of news content. Ultimately, the given news content is classified as true news or fake news often by concatenating and feeding all of these features to a well-trained classifier such as a softmax.

This general procedure can be improved to, for example, facilitate explainable fake news detection and enhance feature representativeness (more discussions on *explainable fake news detection*

المنارة للاستشارات
www



(a) EANN (directly from [Wang et al. 2018]) (b) SAFE (directly from [Zhou et al. 2020])

Fig. 6. Multimodal fake news detection models.

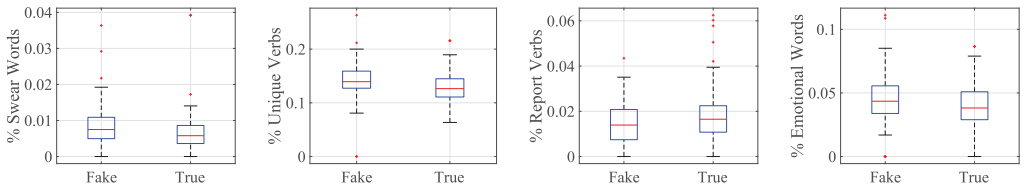


Fig. 7. Fake news textual patterns [Zhou et al. 2019a] (PolitiFact, data is from FakeNewsNet [Shu et al. 2018]). Compared to true news text, fake news text has (i) higher informality (% swear words), (ii) diversity (% unique verbs), and (iii) subjectivity (% report verbs), and is (iv) more emotional (% emotional words).

are provided in Section 6). An example is the event adversarial neural network (EANN) [Wang et al. 2018], which can enhance feature representativeness by extracting features that are invariant under different world events to represent news content (text and images). Figure 6(a) presents the architecture of the EANN model, which has three components: (i) a *multimodal feature extractor*, which extracts both textual and visual features from a given news article using neural networks; (2) an *event discriminator*, which further captures event-invariant features of the given news by playing a *min-max game*; and (iii) a *fake news detector* for news classification (true or fake). In the EANN, the event discriminator ensures that the extracted features are representative. Another example is SAFE, a multimodal fake news detection method (see its architecture in Figure 6(b)). SAFE explores the relationship between the textual and visual features in a news article to detect fake news based on the falsity that can exist in the news multimodal and/or relational information [Zhou et al. 2020]. Specifically, SAFE assumes that a "gap" often exists between the textual and visual information of fake news by observing that (i) to attract public attention, some fake news writers prefer to use attractive while irrelevant images, and (ii) when a fake news article tells a fake story, it is difficult to find both pertinent and non-manipulated images to match such fake contents.

3.3 Patterns of Fake News Content Style

Some patterns of fake news content (text and images) style are distinguishable from those of the true news. Recent studies have revealed such patterns [Jin et al. 2017; Zhou et al. 2019a]. In particular,

- *Fake news text*, compared to true news text as shown in Figure 7, has higher (i) informality (% swear words), (ii) diversity (% unique verbs), and (iii) subjectivity (% report verbs), and is (iv) more emotional (% emotional words).
- *Fake news images*, compared to true news images as illustrated in Figure 8, often have higher clarity and coherence but while lower diversity and clustering scores (see Table 5 for a description of these features).

المنارة للاستشارات
www

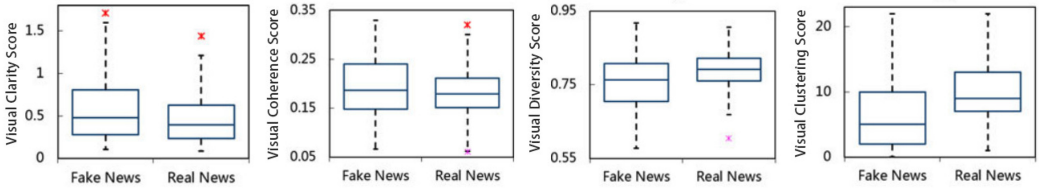


Fig. 8. Fake news visual patterns (Twitter+Weibo, directly from Jin et al. [2017]). Compared to true news images, fake news images often have higher clarity and coherence but lower diversity and clustering score (see Table 5 for a description of these features).

3.4 Discussion

We have detailed how to represent and classify news content style, the two main components of style-based fake news detection, along with some (textual and visual) patterns within fake news content that can help distinguish it from true news content. It should be pointed out that patterns in Figure 7 are limited to political news articles, and those in Figure 8 target a mix of English (Twitter) and Chinese (Weibo) news articles from vague domains. Hence, a more comprehensive analysis of fake news content style across different domains, languages, time periods, and so forth is highly encouraged, as fake news content style may vary across domains and languages, evolve over time, and ultimately impact prediction performance [Pérez-Rosas et al. 2017] (more discussions on this topic are presented in Section 6). Furthermore, the writing style can be manipulated. Style-based (or knowledge-based) fake news detection relies heavily on news content, which enables the models to identify fake news before it has been propagated on social media (i.e., to achieve *fake news early detection*, which we discuss more in Section 6). However, such heavy dependence "helps" malicious entities bypass style-based models by changing their writing style. In other words, style-based fake news detection sometimes can be a cat-and-mouse game; any success at detection, in turn, will inspire future countermeasures by fake news writers. To resist such an attack, one can further involve social context information into news analysis to enhance model robustness, which we will discuss next in Sections 4 and 5.

4 PROPAGATION-BASED FAKE NEWS DETECTION

When detecting fake news from a propagation-based perspective, one can investigate and utilize the information related to the dissemination of fake news, such as how users spread it. Similar to style-based fake news detection, propagation-based fake news detection is often formulated as a binary (or multi-label) classification problem as well, but with a different input. Broadly speaking, the input to a propagation-based method can be either a (I) *news cascade*, a direct representation of news propagation, or a (II) self-defined graph, an indirect representation capturing additional information on news propagation. Hence, propagation-based fake news detection boils down to classifying (I) news cascades or (II) self-defined graphs. We review fake news detection using news cascades in Section 4.1 and fake news detection using self-defined propagation graphs in Section 4.2, and we provide our discussions in Section 4.3.

4.1 Fake News Detection Using News Cascades

We first define a *news cascade* in Definition 8, a formal representation of news dissemination that has been adopted by many studies (e.g., Castillo et al. [2011], Ma et al. [2018], Vosoughi et al. [2018], and Wu et al. [2015]).

Definition 8 (News Cascade). A news cascade is a tree or tree-like structure that directly captures the propagation of a certain news article on a social network (Figure 9 provides examples). The

المنارة للاستشارات
www

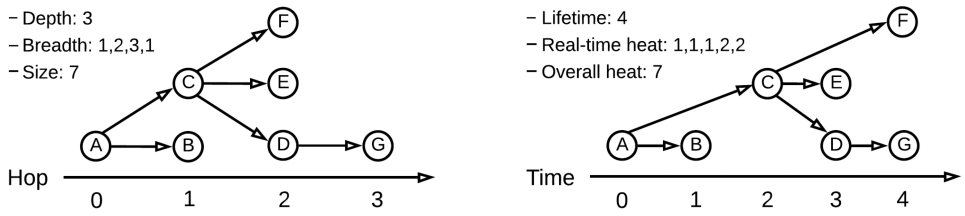


Fig. 9. Illustrations of news cascades.

root node of a news cascade represents the user who first shared the news article (i.e., initiator); other nodes in the cascade represent users that have subsequently spread the article by forwarding it after it was posted by their parent nodes, which they are connected to via edges. A news cascade can be represented in terms of the number of steps (i.e., hops) that the news has traveled (i.e., hop-based news cascade) or the times that it was posted (i.e., time-based news cascade).

Hop-based news cascade, often a standard tree, allowing natural measures such as

—Depth: the maximum number of steps (hops) that the news has traveled within a cascade;
—Breadth (at hop k): the number of users who have spread the news k steps (hops) after it was initially posted within a cascade; and
—Size: the total number of users in a cascade.

Time-based news cascade, often a tree-like structure, allowing natural measures such as

—Lifetime: the longest interval during which the news has been propagated;
—Real-time heat (at time t): the number of users posting/forwarding the news at time t; and
—Overall heat: the total number of users who have forwarded/posted the news.

Note that a specific news article can lead to multiple simultaneous cascades due to multiple initiating users. Furthermore, often within a news cascade, nodes (users) are represented with a series of attributes and additional information, such as whether they (support or oppose) the fake news, their profile information, previous posts, and their comments.

Based on Definition 8, classifying a news article (true or fake) using its cascade(s) boils down to classifying its cascade(s) as true or fake. To perform this classification, some proposed methods rely on (I) traditional ML, whereas others utilize (II) (deep) neural networks.

► *Traditional ML models.* Within a traditional ML framework, to classify a news cascade that has been represented as a set of features, one often relies on supervised learning methods such as SVMs [Castillo et al. 2011; Kwon et al. 2013; Wu et al. 2015], decision trees [Castillo et al. 2011; Kwon et al. 2013], decision rules [Castillo et al. 2011], naive Bayes [Castillo et al. 2011], and RF [Kwon et al. 2013]. Table 7 summarizes the news cascade features used in current research.

As shown in Table 7, cascade features can be inspired by *fake news propagation patterns* observed in empirical studies. For example, a recent study has investigated the differences in diffusion patterns of all verified true and fake news stories on Twitter from 2006 to 2017 [Vosoughi et al. 2018]. The study reveals that fake news spreads faster, farther, and more widely, and is more popular with a higher structure virality score, compared to true news. Specifically, the cascade depth, maximum breadth (and mean breadth at every depth), size, and structural virality [Goel et al. 2015] (i.e., the average distance among all pairs of nodes in a cascade) for fake news are generally greater than that of true news. In addition, the time it takes for fake news cascades to reach any depth (and size) is less than that for true news cascades (Figure 10).

المنارة للاستشارات
www

Table 7. News Cascade Features

Feature Type	Feature Description	H	T	P
Cascade size	Overall number of nodes in a cascade [Castillo et al. 2011; Vosoughi et al. 2018]	✓	✓	✓
Cascade breadth	Maximum (or average) breadth of a news cascade [Vosoughi et al. 2018]	✓		✓
Cascade depth	Depth of a news cascade [Castillo et al. 2011; Vosoughi et al. 2018]	✓		✓
Structural virality	Average distance among all pairs of nodes in a cascade [Vosoughi et al. 2018]	✓		✓
Node degree	Degree of the root node of a news cascade [Castillo et al. 2011]	✓		
	Maximum (or average) degree of non-root nodes in a news cascade [Castillo et al. 2011]	✓		
Spread speed	Time taken for a cascade to reach a certain depth (or size) [Vosoughi et al. 2018]	✓		✓
	Time interval between the root node and its child nodes [Wu et al. 2015]		✓	
Cascade similarity	Similarity scores between a cascade and other cascades in the corpus [Wu et al. 2015]	✓		

H, hop-based news cascades; T, time-based news cascades; P, pattern-driven features.

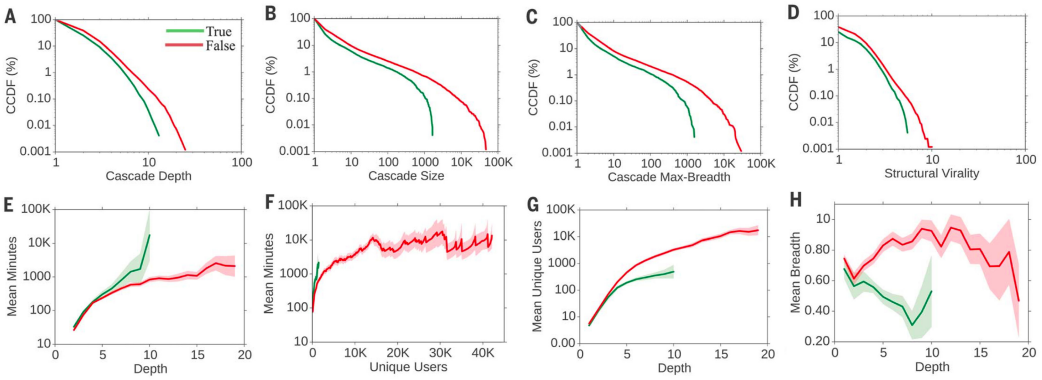
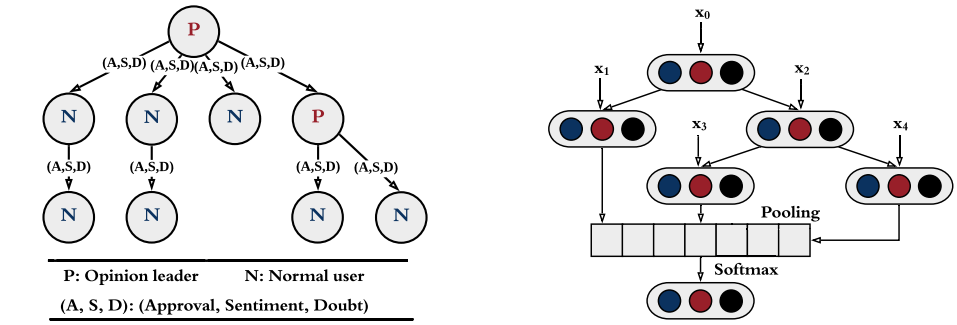


Fig. 10. Fake news cascade-based propagation patterns (Twitter data, directly from Vosoughi et al. [2018]). (A–D) Complementary cumulative distribution function (CCDF) distributions of cascade depth, size, max-breadth, and structural virality of fake news are always above that of true news. (E, F) The average time taken for fake news cascades to reach a certain depth and a certain number of unique users are both less than that for true news cascades. (G, H) For fake news cascades, their average number of unique users and breadth at a certain depth are always greater than that of true news cascades.

The cascade features in Table 7 can be viewed from another perspective, where they can either capture the local structure of a cascade, such as its depth and width [Castillo et al. 2011; Vosoughi et al. 2018], or allow comparing the overall structure of a cascade with that of other cascades by computing similarities [Wu et al. 2015]. A common strategy to compute such graph similarities is to use *graph kernels* [Vishwanathan et al. 2010]. For example, Wu et al. [2015] develop a *random walk graph kernel* to compute the similarity between two hop-based news cascades that contain additional information, such as the user roles (*opinion leader* or *normal user*) and approval, sentiment, and doubt scores for user posts (see the structure of such a cascade in Figure 11(a)).

► *DL models.* Within a DL framework, learning the representation of news cascades often relies on neural networks, where a softmax function often acts as a classifier. For example, Ma et al. develop recursive neural networks (RvNNs), a tree-structured neural network, based on news cascades [Bian et al. 2020; Ma et al. 2018]. A top-down RvNN model with gated recurrent units (GRUs) is shown in Figure 11(b). Specifically, for each node j with a post on a certain news report represented as a TF-IDF vector $\mathbf{x}_j$, its hidden state $\mathbf{h}_j$ is recursively determined by $\mathbf{x}_j$ and the hidden

المنارة للاستشارات
www



(a) A Traditional Machine Learning Model [Wu et al. 2015]

(b) A Deep Learning Model [Ma et al. 2018]

Fig. 11. Examples of cascade-based fake news detection models.

state of its parent node $\mathcal{P}(j)$, denoted as $\mathbf{h}_{\mathcal{P}(j)}$. Formally, $\mathbf{h}_j$ is derived using a standard GRU formulation:

$$\begin{aligned} \mathbf{r}_j &= \sigma(\mathbf{W}_r \mathbf{x}_j \mathbf{V} + \mathbf{U}_r \mathbf{h}_{\mathcal{P}(j)}), \\ \mathbf{z}_j &= \sigma(\mathbf{W}_z \mathbf{x}_j \mathbf{V} + \mathbf{U}_z \mathbf{h}_{\mathcal{P}(j)}), \\ \mathbf{h}_j &= \mathbf{z}_j \odot \tanh(\mathbf{W}_h \mathbf{x}_j \mathbf{V} + \mathbf{U}_h(\mathbf{h}_{\mathcal{P}(j)} \odot \mathbf{r}_j)) + (1 - \mathbf{z}_j) \odot \mathbf{h}_{\mathcal{P}(j)}, \end{aligned} \tag{5}$$

where $\mathbf{r}_j$ is a reset gate vector, $\mathbf{z}_j$ is an update gate vector, $\mathbf{W}_*$, $\mathbf{U}_*$, and $\mathbf{V}$ denote parameter matrices, $\sigma(\cdot)$ is the sigmoid function, $\tanh(\cdot)$ is the hyperbolic tangent, and $\odot$ denotes entry-wise product. In this way, a representation (hidden state) is learned for each leaf node in the cascade. These representations are used as inputs to a *max-pooling* layer, which computes the final representation $\mathbf{h}$ for the news cascade. The max-pooling layer takes the maximum value of each dimension of the hidden state vectors over all of the leaf nodes. Finally, the label of news cascade is predicted as

$$\tilde{y} = \text{Softmax}(\mathbf{Q}\mathbf{h} + \mathbf{b}), \tag{6}$$

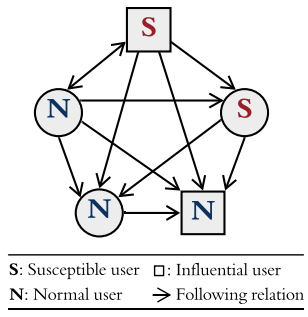
where $\mathbf{Q}$ and $\mathbf{b}$ are parameters. The model (parameters) can be trained (estimated) by minimizing some cost function, such as squared error [Ma et al. 2018] and cross-entropy [Zhang et al. 2018], using some optimization algorithm, such stochastic gradient descent (SGD) [Wang et al. 2018], Adam [Kingma and Ba 2014], and the alternating direction method of multipliers (ADMM) [Boyd et al. 2011; Shu et al. 2019c].

4.2 Fake News Detection Using Self-Defined Propagation Graphs

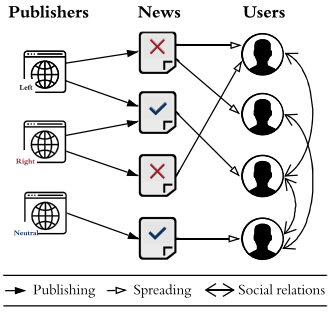
When detecting fake news using self-defined propagation graphs (networks), one constructs flexible networks to indirectly capture fake news propagation. These networks can be homogeneous, heterogeneous, or hierarchical.

▶ *Homogeneous network.* Homogeneous networks are networks containing a single type of node and a single type of edge. One example is the *news spreader network* (Figure 12(a)), a subgraph of social network of users, where each network is for a news article; each node in the network is a user spreading the news; and an edge between two nodes indicates the following relationship of two news spreaders [Zhou and Zafarani 2019]. Classifying a news article (fake or true) using its spreader network is equivalent to classifying the network. Recently, in an earlier work, we analyzed such networks at the level of *node*, *ego*, *triad*, *community*, and the overall *network* [Zhou and Zafarani 2019], respectively, and revealed four patterns within fake news spreader networks: More-Spreader Pattern (i.e., more users spread fake news than true news), Farther-Distance Pattern (i.e., fake news spreads farther than true news), Stronger-Engagement Pattern (i.e., spreaders engage more strongly with fake news than with true news), and Denser-Network

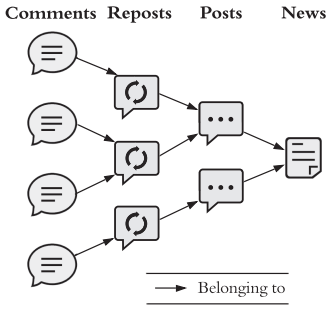
المنارة للاستشارات
www



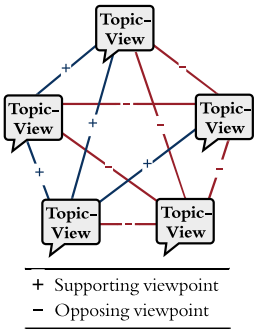
(a) Spreader Net. [Zhou and Zafarani 2019]



(a) [Shu et al. 2019c]

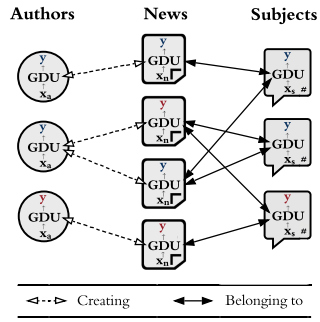


(a) [Shu et al. 2019b]



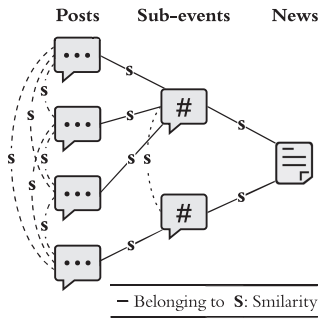
(b) Stance Net. [Jin et al. 2016]

Fig. 12. Homogeneous networks.



(b) [Zhang et al. 2018]

Fig. 13. Heterogeneous networks.



(b) [Jin et al. 2014]

Fig. 14. Hierarchical networks.

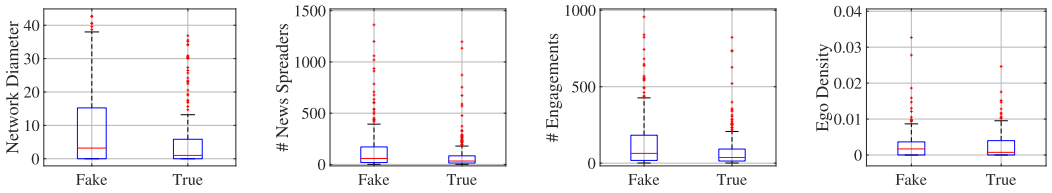
Pattern (i.e., fake news spreaders form denser networks compared to true news spreaders) (see Figure 15). Another example of a homogeneous network is a *stance network* [Jin et al. 2016], where nodes are news-related posts by users and edges represent supporting (+) or opposing (–) relations among each pair of posts, such as the similarity between each pair of posts that can be calculated using a distance measure such as Jensen-Shannon divergence [Jin et al. 2016] or Jaccard distance [Jin et al. 2014]. The stance network is shown in Figure 12(b). Fake news detection using a stance network boils down to evaluating the credibility of news-related posts (i.e., lower credibility = fake news), which can be further cast as a graph optimization problem. Specifically, let $\mathbf{A} \in \mathbb{R}^{n \times n}$ denote the adjacency matrix of the aforementioned stance network with n posts and $\mathbf{c} = (c_1, \ldots, c_n)$ denote the vector of post credibility scores. Assuming that supporting posts have similar credibility scores, the cost function in the work of Zhou et al. [2004a] can be used and the problem can be defined as

$$\arg\min_{\mathbf{c}} \mu||\mathbf{c} - \mathbf{c}_0||^2 + (1-\mu)\sum_{i,j=1}^{n} \mathbf{A}_{ij}\left(\frac{\mathbf{c}_i}{\sqrt{\mathbf{D}_{ii}}} - \frac{\mathbf{c}_j}{\sqrt{\mathbf{D}_{jj}}}\right)^2, \tag{7}$$

where $\mathbf{c}_0$ refers to true credibility scores of training posts, $\mathbf{D}_{ij} = \sum_k \mathbf{A}_{ik}$, and $\mu \in [0, 1]$ is a weight parameter.

► *Heterogeneous network.* Heterogeneous networks have multiple types of nodes or edges. An early instance is a (users)-(tweets)-(news events) network in which Gupta et al. [2012] evaluate news credibility by designing an algorithm similar to PageRank [Page et al. 1999]. Their algorithm

المنارة للاستشارات
www



(a) Farther-Distance (b) More-Spreaders (c) Stronger-Engagements (d) Denser-Networks

Fig. 15. Fake news network-based propagation patterns [Zhou and Zafarani 2019] (PolitiFact+BuzzFeed, data are from FakeNewsNet [Shu et al. 2018]). Compared to the truth, fake news can spread farther (a) and attract more spreaders (b), where these spreaders are often more strongly engaged with fake news (c) and more densely connected within fake news spreader networks (d).

determines credibility relationships among these three entities with the assumption that users with low credibility tend to post tweets that are often related to fake news. Another example is the network capturing relationships among news publishers, news articles, and news spreaders (users) shown in Figure 13(a). Using this network, Shu et al. [2019c] classify news articles (as fake news or true news) using a linear classifier, where

(I) News article are represented using latent features, derived using non-negative matrix factorization (NMF):

$$\min \left\| \mathbf{X} - \mathbf{D}\mathbf{V}^\top \right\|_F^2 \quad \text{s.t. } \mathbf{D}, \mathbf{V} \geq 0, \tag{8}$$

where $\mathbf{X} \in \mathbb{R}_+^{m \times t}$ is the given article-word matrix for m news articles. $\mathbf{X}$ will be factorized as $\mathbf{D} \in \mathbb{R}_+^{m \times d}$ (i.e., article-latent feature matrix) and some weight matrix $\mathbf{V} \in \mathbb{R}_+^{t \times d}$.

(II) Assuming that the political bias for each publisher is known—left (−1), least-biased (0), or right (+1)—let $\mathbf{b} \in \{-1, 0, 1\}^l$ denote the political biases for l publishers. Assuming that the political biases of publishers can be represented by the latent features of their published articles, the publisher-article relationship is derived by

$$\min \|\bar{\mathbf{P}}\mathbf{D}\mathbf{q} - \mathbf{b}\|_2^2, \tag{9}$$

where $\bar{\mathbf{P}} \in \mathbb{R}^{l \times m}$ is the normalized publisher-article relation (adjacency) matrix, and $\mathbf{q} \in \mathbb{R}^d$ is the weights.

(III) Assuming that non-credible (credible) users spread fake (true) news, a similar optimization formulation can be utilized to derive the relationship between spreaders (users) and new articles.

Finally, Zhang et al. [2018] develop a framework for a heterogeneous network of users (news authors), news articles, and news subjects to detect fake news (see Figure 13(b)). They introduce *gated diffusive units* (GDUs), a neural network component that can jointly learn the representations of users, news articles, and news subjects [Zhang et al. 2018].

► *Hierarchical network.* In hierarchical networks, various types of nodes and edges form set-subset relationships (i.e., a *hierarchy*). One example is the news-tweet-retweet-reply network (Figure 14(a)), an extension of the news cascade defined in Definition 8 [Shu et al. 2019b]. Hence, the same features listed for news cascades in Table 7 (e.g., cascade depth and size) can be used to predict fake news using this hierarchical network within a traditional ML framework. Another example of a hierarchical network is shown in Figure 14(b), which includes relationships across (i.e., *hierarchical relationships*) and within (i.e., *homogeneous relationships*) news events, sub-events,

المنارة للاستشارات
www

and posts. In such networks, news verification can be transformed into a graph optimization problem [Jin et al. 2014], extending the optimization in Equation (7).

4.3 Discussion

We have discussed the current solutions for detecting fake news by investigating how news spreads online. By involving dissemination information (i.e., social context) in fake news detection, propagation-based methods are more robust against writing style manipulation by malicious entities. However, propagation-based fake news detection is inefficient for *fake news early detection* (see Section 6 for more details), as it is difficult for propagation-based models to detect fake news before it has been disseminated, or to perform well when limited news dissemination information is available.

Furthermore, mining news propagation and news writing style allow one to assess news intention. As discussed, the intuition is that (i) news created with a malicious intent (i.e., to mislead and deceive the public) aims to be "more persuasive" compared to those not having such aims, and (ii) malicious users often play a part in the propagation of fake news to enhance its social influence [Leibenstein 1950]. However, to evaluate if news intentions are properly assessed, one relies on the ground truth (news labels) in training datasets often annotated by domain experts. This ground truth dependency particularly exists when predicting fake news by (semi-)supervised learning within graph optimization [Shu et al. 2019c], traditional statistical learning [Zhou and Zafarani 2019], or deep neural networks [Zhang et al. 2018]. Most current fake news datasets have not provided a clear-cut declaration on whether the annotations within the datasets consider news intention, or how the annotators have manually evaluated it, which motivates the construction of fake news datasets or repositories that provide information on news intention.

Finally, research has shown that political fake news spreads faster, farther, and more widely, and is more popular with a higher structure virality score, than fake news in other domains such as business, terrorism, science, and entertainment [Vosoughi et al. 2018]. Discovering more dissemination patterns of fake news is hence highly encouraged by comparing it with true news or fake news from different domains and languages. Such patterns can deepen the public understanding of fake news and enhance explainability of fake news detection (we discuss *explainable fake news detection* in Section 6).

5 SOURCE-BASED FAKE NEWS DETECTION

One can detect fake news by assessing the *credibility* of its source, where credibility is often defined in the sense of quality and believability—"offering reasonable grounds for being believed" [Castillo et al. 2011; Viviani and Pasi 2017]. As presented in Figure 1, there are three stages within a (fake) news life cycle: being created, published online, and propagating on social media. This section will present how credibility of news stories can be assessed based on that of their sources at each stage. In other words, we deem "sources" as a general concept that includes (I) the sources creating the news stories, such as the news writers (Section 5.1); (II) the sources that publish the news stories, such as the news publishers (Section 5.1); and (III) the sources that spread the news stories on social media, such as the social media accounts (users, Section 5.2). We combine (I) and (II) into one subsection for two reasons: (i) not many studies have investigated news authors, and (ii) news authors and publishers often form an employee-employer relationship; hence, their credibilities intuitively should have some correlations. Note that although the role of news authors, publishers, and spreading users in detecting fake news has been illustrated in Sections 2 through 4 (see related work such as that of Shu et al. [2019c], Zhang et al. [2018], and our earlier work [Zhou and Zafarani 2019], the focus has been on each news story. In contrast, in this section, the focus is on each author, publisher, and user who might have written, published, or spread the news stories. That is why

المنارة للاستشارات
www

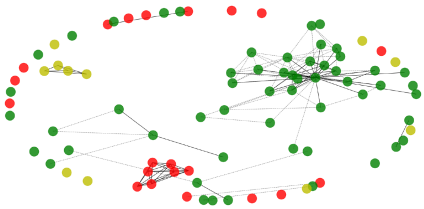


Fig. 16. Coauthorship network (directly from Sitaula et al. [2020]). Nodes are news authors associated with fake news (red), true news (green), or both (yellow), and edges indicate that two authors collaborate only once (dashed) or at least twice (solid).

Fig. 17. Content sharing network (directly from Horne et al. [2019]). Nodes are news publishers, and edges are the flows of news articles among publishers. Orange: Russian/conspiracy community; yellow: right-wing/conspiracy community; green: U.S. mainstream community; magenta: left-wing blog community; and cyan: UK mainstream community.

we, to some extent, regard assessing the credibility of news sources as an indirect way to detect fake news—for instance, one can consider news articles from unreliable news sources as fake news, although it is not unlikely for these sources to post true news. Such an approach to detecting fake news might seem arbitrary but is efficient [Nørregaard et al. 2019], as evidence has revealed that many fake news stories come from either fake news websites that only publish hoaxes or from hyper-partisan websites that present themselves as publishing real news [Silverman 2016].

5.1 Assessing Source Credibility Based on News Authors and Publishers

Based on existing related studies, we organize this section as follows. We first present (I) some patterns shared among (un)reliable authors or publishers. These patterns can help assess the credibility of unknown authors or publishers by exploring their relationships with other authors or publishers. Next, we review (II) web spam detection as a representation of the way to detect unreliable websites. Such methods can be used to identify spam publishers. Finally, we introduce (III) resources that can help obtain the credibility (or political bias) of news publishers (ground truth).

I. Patterns of (un)reliable authors or publishers. Research has shown that news authors and publishers exhibit *homogeneity* in the networks that they form. Specifically, Sitaula et al. [2020] construct the collaboration network of news authors (i.e., *coauthorship network*) based on the FakeNewsNet dataset [Shu et al. 2018]. The network is presented in Figure 16. In this network, each node represents a news author, and the edge between two authors indicates that they collaborate in writing one (the dashed line) or more (the solid line) news articles. All nodes (authors) are grouped as (i) *true-news authors* (green nodes), who are associated with two or more true news stories; (ii) *fake-news authors* (red nodes), who are associated with two or more fake news stories; and (iii) the authors who have published both fake and true news stories (yellow nodes). As we can observe from Figure 16, the coauthorship network exhibits homogeneity: the authors within the same group are more densely connected compared to the authors from different groups.

المنارة للاستشارات
www

Similarly, such homogeneity has been observed in the network formed by news publishers (i.e., *content sharing network*). Figure 17 presents an illustration [Horne et al. 2019]. In this directed network, each node represents a news publisher, and the edge between two publishers indicates the flow of news articles. Nodes (publishers) can be clearly grouped into five communities: Russian/conspiracy community (orange nodes), right-wing/conspiracy community (yellow nodes), U.S. mainstream community (green nodes), left-wing blog community (magenta nodes), and UK mainstream community (cyan nodes), representing various nations and, importantly, various political biases.

II. Web (publisher) spam detection. News publishers often publish their news articles on their own websites. Detecting unreliable publishers thus can be reduced to detecting low-credible websites. To assess website credibility, many practical techniques have been developed, such as web ranking algorithms. Traditional web ranking algorithms such as PageRank [Page et al. 1999] and HITS [Kleinberg 1999] assess website credibility to improve search engine responses to user search queries. However, the weaknesses of these web ranking algorithms provide opportunities for *web spam*, to try to unjustifiably improve website rankings, thus motivating the development of *web spam detection*. A comprehensive survey is available in the work of Spirin and Han [2012]. Web spam can be categorized as (i) *content spam*, which leads to a spam webpage appearing among normal search results primarily due to fake word frequencies (e.g., TF-IDF scores; content spam includes spamming of title, body, meta-tags, anchors, and URLs); (ii) (outgoing and incoming) *link spam*, where the former mostly targets algorithms similar to HITS to achieve high hub scores, and the latter enhances website authority scores by attacking algorithms similar to PageRank; and (iii) other types of spam, such as *cloaking*, *redirection*, and *click spams*. Algorithms to identify Web spam thus can be classified into (i) content-based algorithms, which analyze web content features, such as word counts and content duplication [Baly et al. 2018; Fetterly et al. 2005; Ntoulas et al. 2006]; (ii) link-based algorithms, which detect web spam by utilizing graph information [Horne et al. 2019; Zhou and Pei 2009], learning statistical anomalies [Dong et al. 2015], and performing techniques such as (dis)trust propagation [Gyöngyi et al. 2004], link pruning [Bharat and Henzinger 1998], and graph regularization [Abernethy et al. 2010]; and (iii) other algorithms that are often based on click streams [Dou et al. 2008] or user behavior [Liu et al. 2015].

III. Resources for understanding news publishers. We introduce several resources that can help obtain the ground truth on the credibility (or political bias) of news publishers. One resource is the *Media Bias/Fact Check* website,[19] which provides a list of media along with their political slant: left, left-center, least biased, right-center, and right. Another source is NewsGuard,[18] which relies on expert-based evaluations and rates the reliability of a news source in terms of nine criteria: if it repeatedly publishes false content, responsibly presents information, regularly corrects or clarifies errors, responsibly handles the difference between news and opinion, avoids deceptive headlines, discloses ownership and financing, clearly labels advertising, reveals who is in charge that includes any possible conflicts of interest, and provides information about content creators. Finally, a system named *MediaRank* ranks news sources associated with their peer reputation, reporting bias, bottomline financial pressure, and popularity [Ye and Skiena 2019].

5.2 Assessing Source Credibility Based on Social Media Users

Social media users can be the initiating source for a news story spreading on social media. Intuitively, users with low credibility are more likely to become the spreading source of a fake news

[19] https://mediabiasfactcheck.com/.

المنارة للاستشارات
www

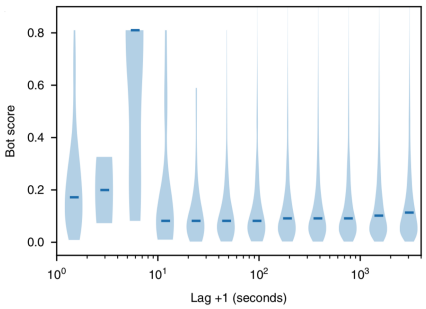


Fig. 18. Temporal engagements of social bots (directly from Shao et al. [2018]), which indicates that bots spread unreliable news at an earlier time compared to humans.

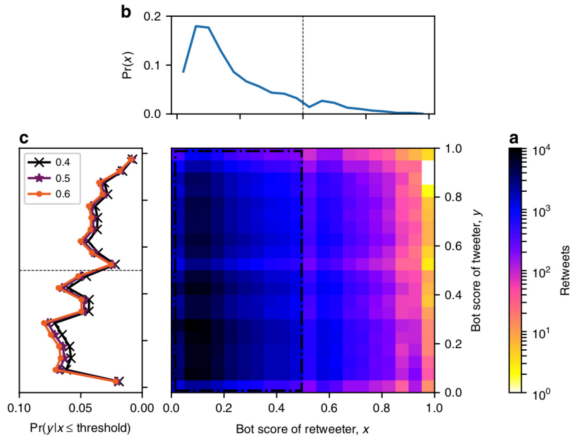


Fig. 19. Impacts of bots on humans (directly from Shao et al. [2018]). (a) Joint distribution of bot scores for tweeters and retweeters of unreliable news. (b) Distribution of bot scores for retweeters, who are mostly humans. (c) Distribution of bot scores for tweeters, where bots and humans both share a significant proportion.

story than reliable users. Here we define and group such low-credible users as (I) malicious users and (II) normal users who are vulnerable to spreading fake news (or any disinformation).

I. Identifying malicious users. Identifying malicious users can often be reduced to detecting *social bots*, a software application that runs automated tasks (scripts) over the Internet. Note that social bots can be benign and designed to provide useful services, such as automatically aggregating content from different sources, like simple news feeds.

Meanwhile, some bots are created to harm by manipulating social media discourse and deceiving social media users, for example, by spreading fake news [Ferrara et al. 2016]. One recent study—that classifies accounts based on observable features such as sharing behavior, the number of social ties, and linguistic features—estimates that between 9% and 15% of active Twitter accounts are bots [Varol et al. 2017]. It has been suggested that millions of social bots have participated in online discussions around the 2016 U.S. presidential election,[20] and some of the same bots were later used to attempt to influence the 2017 French election [Ferrara 2017]. Research has significantly contributed to differentiating bots from humans on social media. For example, a feature-based system, Botometer (formally BotOrNot),[21] extracts six main groups of features (*network*, *user*, *friend*, *temporal*, *content*, *sentiment*) of a Twitter account and uses RF to predict whether it is a bot or not with a 0.95 area under the receiver operating characteristic curve performance. Similarly, Cai et al. [2017] propose a bot detection model that uses DL to learn features from both user posts and behavior. Morstatter et al. [2016] work on striking the balance between precision and recall value in predicting social bots. From 14 million messages spreading 400,000 articles on Twitter during a 10-month period in 2016 and 2017, Shao et al. [2018] find evidence that bots spread unreliable news at an earlier time compared to humans to increase the chances that an article goes "viral" (Figure 18). They also discover that humans do most of the retweeting, and they retweet articles posted by (likely) bots almost as much as those by other humans (Figure 19). Such discoveries demonstrate

[20] http://comprop.oii.ox.ac.uk/research/public-scholarship/resource-for-understanding-political-bots/.
[21] https://botometer.iuni.iu.edu/.

المنارة للاستشارات
www

the difficulty in recognizing social bots and the vulnerability of humans (or normal users) to online information with low credibility. Next, we will discuss the possible ways to identify this group of vulnerable normal users, which, thus far, has been rarely investigated.

II. Identifying vulnerable normal users. Fake news, unlike information such as fake reviews [Jindal and Liu 2008], can "attract" both malicious and normal users, where each group is attracted for different reasons (i.e., different intentions). Malicious users such as some social bots [Ferrara et al. 2016]) often spread fake news intentionally. In contrast, some normal users can frequently and unintentionally do so without recognizing the falsehood. Compared to normal users who are "immune" to fake news, these normal users are "susceptible" and vulnerable to fake news—that is, their credibility is relatively lower compared to other normal users. Theories in Table 2 imply that a user engages in spreading fake news less intentionally when spreading bears a greater (i) *social influence* (e.g., more users spreading the news) [Ashforth and Mael 1989; Boehm 1994; Deutsch and Gerard 1955; Kuran and Sunstein 1999; Leibenstein 1950; MacLeod et al. 1986] and (ii) *self-influence* (e.g., high similarity between the knowledge within the fake news and the preexisting knowledge of the user) [Fisher 1993; Freedman and Sears 1965; Nickerson 1998]. Preexisting knowledge can be captured by user posts. Nevertheless, few studies have quantified the impact of social and self-influence on the vulnerability of users to fake news, or have considered the effect of intention in assessing user credibility and further in fake news intervention (see Section 6 for more discussions on *fake news intervention*).

6 DISCUSSION AND FUTURE WORK

We detailed four fake news detection strategies (*knowledge*-based, *style*-based, *propagation*-based, and *source*-based) separately in Sections 2 through 5; however, they are not independent. Predicting fake news jointly from multiple perspectives is encouraged, where one can combine their strengths. Furthermore, there is a motivation to detail open issues shared among different strategies. Based on fake news characteristics and the current state of fake news research, we highlight the following potential research tasks that can facilitate a deeper understanding of fake news, and improve the interpretability and performance of current fake news detection studies.

I. Detection of non-traditional fake news. Based on how fake news was defined, fake news can also be news articles with (i) outdated knowledge (e.g., "Britain has control over fifty-six colonial countries") or (ii) false claims in *some* parts of the news content (text and images; i.e., news content is partially (in)correct). These non-traditional forms of fake news emphasize different aspects of detection and motivate one to develop more thorough and comprehensive detection strategies. For example, to detect fake news with outdated knowledge, one is encouraged to construct a dynamic KG; to detect fake news that is only partially correct, extending fake news detection to a multi-label classification or regression problem may be more appropriate than defining it as a binary classification problem.

II. Fake news early detection. Fake news early detection aims to detect fake news at an early stage before it becomes widespread so that one can take early actions for fake news mitigation and intervention. Early detection is especially crucial for fake news, as the more fake news spreads, the more likely for people to trust it (i.e., *validity effect* [Boehm 1994]). To detect fake news at an early stage, one has to primarily and efficiently rely on news content and limited social context information, which leads to facing multiple challenges. First, newly emerged events often generate new and unexpected knowledge that has not been stored in existing KGs, or is difficult to be inferred. Second, features that have well represented the style of fake news in one domain or in the past may not be as useful in the other domains or in the future, especially due to the different patterns of fake news in different domains, and the constant evolution of deceptive writing style [Castelo

المنارة للاستشارات
www

et al. 2019]. Finally, limited information may adversely impact the performance of ML methods. To address these challenges and detect fake news early, one can focus on

(1) *timeliness of ground truth*, for example, technologies related to dynamic (real-time) KG construction should be developed to facilitate timely updates of the ground truth;
(2) *feature compatibility* [Wang et al. 2018], specifically, features that can capture the commonalities of deceptive writing style across topics, domains, languages, and the like, as well as the evolution of deceptive writing style; and
(3) *verification efficiency* [Liu and Wu 2018; Zhou et al. 2019a], such as by improving the efficiency of fake news detection by effectively using a small amount of information in a news article (e.g., headlines), or by identifying *check-worthy content* and topics, which we will discuss next.

III. Identifying check-worthy content. With new information created and circulated online at an unprecedented rate, identifying check-worthy content or topics can improve the efficiency of fake news detection and intervention by prioritizing content or topics that are check-worthy [Hassan et al. 2017a]. Whether a given news content or topic is check-worthy can be measured by, for example, (i) its *newsworthiness* or potential to influence the society, or (ii) its historical likelihood of being fake news. Thus, a content or topic that is newsworthy and generally favored by fake news creators is more check-worthy. Newsworthiness can be assessed in many ways, such as by verifying if the title of the news article is a clickbait [Zhou et al. 2019a], if its content will trigger widespread discussions on social media [Vosoughi et al. 2018], or if its topic relates to national affairs and matches with public concerns [Hassan et al. 2017a]. As shown in Figure 2(a), the historical likelihood of a topic being fake news is available in some online resources. For example, "the PolitiFact scorecard"[3] gives statistics on the authenticity distribution of all of the statements on a specific topic [Zhang et al. 2018].

Furthermore, identifying check-worthy portions within a news content (e.g., sentences) is a pathway to *explainable fake news detection*, which we will also discuss as a potential research task in this section. A news portion that contributes more to detecting fake news is more check-worthy; thus, an *attention mechanism* becomes a natural choice for identifying (i.e., weighting) such portions in an RNN-based fake news detection model [Shu et al. 2019a]. Expert-based analyses and justifications provided for each verified news article on fact-checking websites (detailed in Section 2.1) can be potentially combined with such identified portions to provide explanations; however, to date, such research directions have not been well explored.

IV. Cross-domain (-topic, -website, -language) fake news analysis. We highlight this potential research task for two reasons. First, current fake news studies have an emphasis on distinguishing fake news from truth with experimental settings that are generally limited to a particular social network and a specific language. Second, analyzing fake news across domains, topics, websites, and languages allows one to gain a deeper understanding of fake news and identify its unique non-varying characteristics, which can further assist in fake news early detection and in the identification of check-worthy content, both of which we have discussed in this section.

V. Explainable fake news detection. Facilitating interpretability has been of great interest in artificial intelligence [Vilone and Longo 2020] and ML [Du et al. 2018] research. For fake news detection models, interpretability can be provided by mining social feedback, such as the stance taken within (re-)posts and comments [Shu et al. 2019a], and mining expert analyses available on fact-checking websites, yet both have been rarely utilized. Model interpretability can also be provided and enhanced by conducting interdisciplinary research, such as by relying on fundamental theories identified in Table 2 (see Section 1.2). Although there have been studies that have conducted theory- or

المنارة للاستشارات
www

pattern-driven feature engineering to detect fake news within a traditional ML framework [Zhou et al. 2019a; Zhou and Zafarani 2019], there has been limited research on using related theories or domain knowledge to guide the learning process in ML, such as in [deep] neural networks.

VI. Fake news intervention. Fake news studies have emphasized the importance of new business models adopted by social media sites to address fake news intervention, which suggests shifting the emphasis from maximizing user engagement to that on increasing information quality, such as by using self- or government regulations [Lazer et al. 2018]. In addition to introducing new policies and regulations, efficiently blocking and mitigating the spread of fake news also demands technical innovations and developments. Technically, a fake news intervention strategy can be based on network structure, or based on users as we discussed in Section 1.2. When intervening based on network structure, one aims to stop fake news from spreading by blocking its propagation paths, relying on analyzing the network structure of its propagation and predicting how fake news is going to further spread. From a user perspective, fake news intervention relies on specific roles users play in fake news dissemination. One such role is being an *influential user* (i.e., *opinion leader*). When blocking certain fake news in a social network, targeting these influential spreaders leads to a more efficient intervention compared to those with a negligible social influence on others. Another beneficial role is being a *corrector*, users on social networks who take an active role in mitigating the spread of fake news by attaching links to their posts or comments that debunk the fake news [Vo and Lee 2018]. Furthermore, the intervention strategy for *malicious users* and *normal users* should be different, as they can both spread fake news; malicious users should be penalized, whereas normal users should be assisted to improve their ability to distinguish fake news. For example, *personal recommendation* of true news articles and/or articles with refuting evidence can be helpful to normal users. These recommendation should not only cater to the topics that the users want to read but also those topics that they are most gullible to due to their political biases or preexisting knowledge.

7 CONCLUSION

This survey extensively reviews and evaluates current fake news research by (I) defining fake news, differentiating it from deceptive news, false news, satire news, misinformation, disinformation, clickbaits, cheery-picking, and rumors based on three characteristics: authenticity, intention, and being news; (II) detailing interdisciplinary fake news research by firstly and comprehensively identifying related fundamental theories in, for example, social sciences; (III) reviewing the methods that detect fake news from four perspectives: the false *knowledge* fake news communicates, its writing *style*, its *propagation* patterns, and the credibility of its *source*; and (IV) highlighting challenges in current research and some research opportunities that go with these challenges. As fake news research is evolving, we accompany this survey with the online repository http://fake-news.site, which will provide summaries and timely updates on the research developments on fake news, including tutorials, recent publications and methods, datasets, and other related resources.

REFERENCES

Jacob Abernethy, Olivier Chapelle, and Carlos Castillo. 2010. Graph regularization methods for web spam detection. *Machine Learning* 81, 2 (2010), 207–225.

Sadia Afroz, Michael Brennan, and Rachel Greenstadt. 2012. Detecting hoaxes, frauds, and deception in writing style online. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'12)*. IEEE, Los Alamitos, CA, 461–475.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31, 2 (2017), 211–236.

Yasser Altowim, Dmitri V. Kalashnikov, and Sharad Mehrotra. 2014. Progressive approach to relational entity resolution. *Proceedings of the VLDB Endowment* 7, 11 (2014), 999–1010.

المنارة للاستشارات

www

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *Proceedings of the 5th International Conference on Learning Representations (ICLR'17)*.

Blake E. Ashforth and Fred Mael. 1989. Social identity theory and the organization. *Academy of Management Review* 14, 1 (1989), 20–39.

Abolfazl Asudeh, H. V. Jagadish, You Will Wu, and Cong Yu. 2020. On detecting cherry-picked trendlines. *Proceedings of the VLDB Endowment* 13, 6 (2020), 939–952.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A nucleus for a web of open data. In *The Semantic Web*. Springer, 722–735.

Péter Bálint and Géza Bálint. 2009. The Semmelweis-reflex. *Orvosi Hetilap* 150, 30 (2009), 1430.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. *arXiv:1810.01765*.

Sudipta Basu. 1997. The conservatism principle and the asymmetric timeliness of earnings. *Journal of Accounting and Economics* 24, 1 (1997), 3–37.

Krishna Bharat and Monika R. Henzinger. 1998. Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, 104–111.

Indrajit Bhattacharya and Lise Getoor. 2007. Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data* 1, 1 (2007), 5.

Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. *arXiv:2001.06362*.

Lawrence E. Boehm. 1994. The validity effect: A search for mediating variables. *Personality and Social Psychology Bulletin* 20, 3 (1994), 285–293.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*. ACM, New York, NY, 1247–1250.

Gary D. Bond, Rebecka D. Holman, Jamie-Ann L. Eggert, Lassiter F. Speller, Olivia N. Garcia, Sasha C. Mejia, Kohlby W. Mcinnes, Eleny C. Ceniceros, and Rebecca Rustige. 2017. 'Lyin' Ted','Crooked Hillary', and 'Deceptive Donald': Language of lies in the 2016 US presidential debates. *Applied Cognitive Psychology* 31, 6 (2017), 668–677.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*. 2787–2795.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning* 3, 1 (2011), 1–122.

Chiyu Cai, Linjing Li, and Daniel Zeng. 2017. Detecting social bots by jointly modeling deep behavior and content information. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, New York, NY, 1995–1998.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2010. Toward an architecture for never-ending language learning. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAA'10)*, Vol. 5. 3.

Sonia Castelo, Thais Almeida, Anas Elghafari, Aécio Santos, Kien Pham, Eduardo Nakamura, and Juliana Freire. 2019. A topic-agnostic approach for identifying fake news pages. In *Companion Proceedings of the 2019 World Wide Web Conference*. ACM, New York, NY, 975–980.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*. ACM, New York, NY, 675–684.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, 785–794.

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

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. arXiv:1406.1078.

Peter Christen. 2008. Automatic record linkage using seeded nearest neighbour and support vector machine classification. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, 151–159.

Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M. Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. *PloS One* 10, 6 (2015), e0128193.

المنارة للاستشارات

www

Sarah Cohen, James T. Hamilton, and Fred Turner. 2011. Computational journalism. *Communications of the ACM* 54, 10 (2011), 66–71.

Niall J. Conroy, Victoria L. Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology* 52, 1 (2015), 1–4.

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*. 273–274.

Yong Deng. 2015. Generalized evidence theory. *Applied Intelligence* 43, 3 (2015), 530–543.

Morton Deutsch and Harold B. Gerard. 1955. A study of normative and informational social influences upon individual judgment.*Journal of Abnormal and Social Psychology* 51, 3 (1955), 629.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805.

Shimin Di, Yanyan Shen, and Lei Chen. 2019. Relation extraction via domain-aware transfer learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, 1348–1357.

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge Vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, 601–610.

Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. 2015. Knowledge-based trust: Estimating the trustworthiness of web sources. *Proceedings of the VLDB Endowment* 8, 9 (2015), 938–949.

Zhicheng Dou, Ruihua Song, Xiaojie Yuan, and JiRong Wen. 2008. Are click-through data adequate for learning web search rankings? In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*. ACM, New York, NY, 73–82.

Mengnan Du, Ninghao Liu, and Xia Hu. 2018. Techniques for interpretable machine learning. arXiv:1808.00033.

David Dunning, Dale W. Griffin, James D. Milojkovic, and Lee Ross. 1990. The overconfidence effect in social prediction.*Journal of Personality and Social Psychology* 58, 4 (1990), 568.

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 (Volume 2: Short Papers)*. 171–175.

Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. 2014. Do we need hundreds of classifiers to solve real world classification problems. *Journal of Machine Learning Research* 15, 1 (2014), 3133–3181.

Emilio Ferrara. 2017. Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday* 22, 8 (2017). Available at http://firstmonday.org/ojs/index.php/fm/article/view/8005/6516.

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

William Ferreira and Andreas Vlachos. 2016. Emergent: A novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1163–1168.

Dennis Fetterly, Mark Manasse, and Marc Najork. 2005. Detecting phrase-level duplication on the World Wide Web. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, 170–177.

Robert J. Fisher. 1993. Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research* 20, 2 (1993), 303–315.

Jonathan L. Freedman and David O. Sears. 1965. Selective exposure. In *Advances in Experimental Social Psychology*. Vol. 2. Elsevier, 57–97.

Nico H. Frijda. 1986. *The Emotions*. Cambridge University Press.

Christie M. Fuller, David P. Biros, and Rick L. Wilson. 2009. Decision support for determining veracity via linguistic-based cues. *Decision Support Systems* 46, 3 (2009), 695–703.

Lise Getoor and Ashwin Machanavajjhala. 2012. Entity resolution: Theory, practice & open challenges. *Proceedings of the VLDB Endowment* 5, 12 (2012), 2018–2019.

Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J. Watts. 2015. The structural virality of online diffusion. *Management Science* 62, 1 (2015), 180–196.

Jennifer Golbeck, Matthew Mauriello, Brooke Auxier, Keval H. Bhanushali, Christopher Bonk, Mohamed Amine Bouzaghrane, Cody Buntain, et al. 2018. Fake news vs satire: A dataset and analysis. In *Proceedings of the 10th ACM Conference on Web Science*. ACM, New York, NY, 17–21.

Manish Gupta, Peixiang Zhao, and Jiawei Han. 2012. Evaluating event credibility on Twitter. In *Proceedings of the 2012 SIAM International Conference on Data Mining*. 153–164.

Zoltán Gyöngyi, Hector Garcia-Molina, and Jan Pedersen. 2004. Combating web spam with TrustRank. In *Proceedings of the 30th International Conference on Very Large Data Bases—Volume 30*. 576–587.

المنارة للاستشارات

www

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017a. Toward automated fact-checking: Detecting check-worthy factual claims by ClaimBuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, 1803–1812.

Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, et al. 2017b. ClaimBuster: The first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment* 10, 12 (2017), 1945–1948.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. 2013. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence* 194 (2013), 28–61.

Benjamin D. Horne, Jeppe Nørregaard, and Sibel Adalı. 2019. Different spirals of sameness: A study of content sharing in mainstream and alternative media. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 257–266.

Carl I. Hovland, O. J. Harvey, and Muzafer Sherif. 1957. Assimilation and contrast effects in reactions to communication and attitude change.*Journal of Abnormal and Social Psychology* 55, 2 (1957), 244.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4700–4708.

Kathleen Hall Jamieson and Joseph N. Cappella. 2008. *Echo Chamber: Rush Limbaugh and the Conservative Media Establishment*. Oxford University Press.

Zhiwei Jin, Juan Cao, Yu-Gang Jiang, and Yongdong Zhang. 2014. News credibility evaluation on microblog with a hierarchical propagation model. In *Proceedings of the IEEE International Conference on Data Mining (ICDM'14)*. IEEE, Los Alamitos, CA, 230–239.

Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. 2016. News verification by exploiting conflicting social viewpoints in microblogs. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI'16)*. 2972–2978.

Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. 2017. Novel visual and statistical image features for microblogs news verification. *IEEE Transactions on Multimedia* 19, 3 (2017), 598–608.

Nitin Jindal and Bing Liu. 2008. Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*. ACM, New York, NY, 219–230.

Marcia K. Johnson and Carol L. Raye. 1981. Reality monitoring. *Psychological Review* 88, 1 (1981), 67.

Edward E. Jones and Daniel McGillis. 1976. Correspondent inferences and the attribution cube: A comparative reappraisal. *New Directions in Attribution Research* 1 (1976), 389–420.

Daniel Kahneman and Amos Tversky. 2013. Prospect theory: An analysis of decision under risk. In *Handbook of the Fundamentals of Financial Decision Making: Part I*. World Scientific, 99–127.

Bingyi Kang and Yong Deng. 2019. The maximum Deng entropy. *IEEE Access* 7, 1 (2019), 120758–120765.

Hamid Karimi and Jiliang Tang. 2019. Learning hierarchical discourse-level structure for fake news detection. arXiv:1903.07389.

Seyed Mehran Kazemi and David Poole. 2018. Simple embedding for link prediction in knowledge graphs. In *Advances in Neural Information Processing Systems*. 4284–4295.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. arXiv:1408.5882.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv:1412.6980.

Jon M. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM* 46, 5 (1999), 604–632.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. arXiv:1806.03713.

Sotiris B. Kotsiantis, I. Zaharakis, and P. Pintelas. 2007. Supervised machine learning: A review of classification techniques. *Emerging Artificial Intelligence Applications in Computer Engineering* 160 (2007), 3–24.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. 1097–1105.

Nir Kshetri and Jeffrey Voas. 2017. The economics of "fake news." *IT Professional*6 (2017), 8–12.

Adam Kucharski. 2016. Post-truth: Study epidemiology of fake news. *Nature* 540, 7634 (2016), 525.

Timur Kuran and Cass R. Sunstein. 1999. Availability cascades and risk regulation. *Stanford Law Review* 51, 4 (1999), 683–768.

Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. Prominent features of rumor propagation in online social media. In *Proceedings of the International Conference on Data Mining*. IEEE, Los Alamitos, CA.

Ni Lao and William W. Cohen. 2010. Relational retrieval using a combination of path-constrained random walks. *Machine Learning* 81, 1 (2010), 53–67.

المنارة للاستشارات

www

David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the International Conference on Machine Learning*. 1188–1196.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436.

Yann LeCun, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard, and Lawrence D. Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural Computation* 1, 4 (1989), 541–551.

Harvey Leibenstein. 1950. Bandwagon, snob, and Veblen effects in the theory of consumers' demand. *Quarterly Journal of Economics* 64, 2 (1950), 183–207.

Hongtao Lin, Jun Yan, Meng Qu, and Xiang Ren. 2019. Learning dual retrieval module for semi-supervised relation extraction. In *Proceedings of the World Wide Web Conference*. ACM, New York, NY, 1073–1083.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*.

Xin Liu, Radoslaw Nielek, Paulina Adamska, Adam Wierzbicki, and Karl Aberer. 2015. Towards a highly effective and robust web credibility evaluation system. *Decision Support Systems* 79 (2015), 99–108.

Yang Liu and Yi-Fang Brook Wu. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18)*.

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)*. 1980–1989.

Colin MacLeod, Andrew Mathews, and Philip Tata. 1986. Attentional bias in emotional disorders.*Journal of Abnormal Psychology* 95, 1 (1986), 15.

Steven A. McCornack, Kelly Morrison, Jihyun Esther Paik, Amy M. Wisner, and Xun Zhu. 2014. Information manipulation theory 2: A propositional theory of deceptive discourse production. *Journal of Language and Social Psychology* 33, 4 (2014), 348–377.

Miriam J. Metzger, Ethan H. Hartsell, and Andrew J. Flanagin. 2015. Cognitive dissonance or credibility? A comparison of two theoretical explanations for selective exposure to partisan news. *Communication Research* (2015), 0093650215613136.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv:1301.3781.

Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. arXiv:1712.00547.

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

Fred Morstatter, Liang Wu, Tahora H. Nazer, Kathleen M. Carley, and Huan Liu. 2016. A new approach to bot detection: Striking the balance between precision and recall. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE, Los Alamitos, CA, 533–540.

Ndapandula Nakashole, Gerhard Weikum, and Fabian Suchanek. 2012. PATTY: A taxonomy of relational patterns with semantic types. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 1135–1145.

Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2016. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE* 104, 1 (2016), 11–33.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2012. Factorizing YAGO: Scalable machine learning for linked data. In *Proceedings of the 21st International Conference on World Wide Web*. ACM, New York, NY, 271–280.

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

Feng Niu, Ce Zhang, Christopher Ré, and Jude Shavlik. 2012. Elementary: Large-scale knowledge-base construction via machine learning and statistical inference. *International Journal on Semantic Web and Information Systems* 8, 3 (2012), 42–73.

Jeppe Nørregaard, Benjamin D. Horne, and Sibel Adalı. 2019. NELA-GT-2018: A large multi-labelled news dataset for the study of misinformation in news articles. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 630–638.

Alexandros Ntoulas, Marc Najork, Mark Manasse, and Dennis Fetterly. 2006. Detecting spam web pages through content analysis. In *Proceedings of the 15th International Conference on World Wide Web*. ACM, New York, NY, 83–92.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2 (2019), 13.

المنارة للاستشارات

www

Ray Oshikawa, Jing Qian, and William Yang Wang. 2018. A survey on natural language processing for fake news detection. arXiv:1811.00770.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report. Stanford InfoLab.

Gabriella Pasi, Marco Viviani, and Alexandre Carton. 2019. A multi-criteria decision making approach based on the Choquet integral for assessing the credibility of user-generated content. *Information Sciences* 503 (2019), 574–588.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. 1532–1543.

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

David Pogue. 2017. How to stamp out fake news. *Scientific American* 316, 2 (2017), 24.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. arXiv:1702.05638.

Emily Pronin, Justin Kruger, Kenneth Savtisky, and Lee Ross. 2001. You don't know me, but I know you: The illusion of asymmetric insight. *Journal of Personality and Social Psychology* 81, 4 (2001), 639.

K. Rapoza. 2017. Can "fake news" impact the stock market? *Forbes*. Retrieved August 13, 2020 from https://www.forbes.com/sites/kenrapoza/2017/02/26/can-fake-news-impact-the-stock-market/#4b3542d52fac.

Paul Resnick, Samuel Carton, Souneil Park, Yuncheng Shen, and Nicole Zeffer. 2014. RumorLens: A system for analyzing the impact of rumors and corrections in social media. In *Proceedings of the Computational Journalism Conference*, Vol. 5.

Victoria L. Rubin. 2010. On deception and deception detection: Content analysis of computer-mediated stated beliefs. *Proceedings of the Association for Information Science and Technology* 47, 1 (2010), 1–10.

Victoria L. Rubin, Yimin Chen, and Niall J. Conroy. 2015. Deception detection for news: Three types of fakes. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*. 83.

Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45, 11 (1997), 2673–2681.

Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th International Conference Companion on World Wide Web*. 745–750.

Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nature Communications* 9, 1 (2018), 4787.

Baoxu Shi and Tim Weninger. 2016. Discriminative predicate path mining for fact checking in knowledge graphs. *Knowledge-Based Systems* 104 (2016), 123–133.

Baoxu Shi and Tim Weninger. 2018. Open-world knowledge graph completion. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019a. dEFEND: Explainable fake news detection. In *Proceedings of the 26th SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'19)*.

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:1809.01286.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, and Huan Liu. 2019b. Hierarchical propagation networks for fake news detection: Investigation and exploitation. arXiv:1903.09196.

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

Kai Shu, Suhang Wang, and Huan Liu. 2019c. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining*. ACM, New York, NY, 312–320.

Michael Siering, Jascha-Alexander Koch, and Amit V. Deokar. 2016. Detecting fraudulent behavior on crowdfunding platforms: The role of linguistic and content-based cues in static and dynamic contexts. *Journal of Management Information Systems* 33, 2 (2016), 421–455.

Craig Silverman. 2016. This analysis shows how viral fake election news stories outperformed real news on Facebook. *BuzzFeed News* 16 (2016). Available at https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.

Niraj Sitaula, Chilukuri K. Mohan, Jennifer Grygiel, Xinyi Zhou, and Reza Zafarani. 2020. Credibility-based fake news detection. In *Disinformation, Misinformation and Fake News in Social Media: Emerging Research Challenges and Opportunities*. Springer.

المنارة للاستشارات

www

Alexander Smith and Vladimir Banic. 2016. Fake news: How a partying Macedonian teen earns thousands publishing lies. *NBC News* 9 (2016). Available at https://www.nbcnews.com/news/world/fake-news-how-partying-macedonian-teen-earns-thousands-publishing-lies-n692451.

Richard Socher, Danqi Chen, Christopher D. Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems*. 926–934.

Nikita Spirin and Jiawei Han. 2012. Survey on web spam detection: Principles and algorithms. *ACM SIGKDD Explorations Newsletter* 13, 2 (2012), 50–64.

Rebecca C. Steorts, Rob Hall, and Stephen E. Fienberg. 2016. A Bayesian approach to graphical record linkage and deduplication. *Journal of the American Statistical Association* 111, 516 (2016), 1660–1672.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. YAGO: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*. ACM, New York, NY, 697–706.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–9.

Edson C. Tandoc Jr., Zheng Wei Lim, and Richard Ling. 2018. Defining "fake news": A typology of scholarly definitions. *Digital Journalism* 6, 2 (2018), 137–153.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: A large-scale dataset for fact extraction and verification. arXiv:1803.05355.

Rakshit Trivedi, Bunyamin Sisman, Xin Luna Dong, Christos Faloutsos, Jun Ma, and Hongyuan Zha. 2018. LinkNBed: Multi-graph representation learning with entity linkage. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 252–262.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *Proceedings of the International Conference on Machine Learning*. 2071–2080.

Udo Undeutsch. 1967. Beurteilung der glaubhaftigkeit von aussagen. *Handbuch der Psychologie* 11 (1967), 26–181.

Onur Varol, Emilio Ferrara, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

Giulia Vilone and Luca Longo. 2020. Explainable artificial intelligence: A systematic review. arXiv:2006.00093.

S. Vichy N. Vishwanathan, Nicol N. Schraudolph, Risi Kondor, and Karsten M. Borgwardt. 2010. Graph kernels. *Journal of Machine Learning Research* 11 (April 2010), 1201–1242.

Marco Viviani and Gabriella Pasi. 2016. A multi-criteria decision making approach for the assessment of information credibility in social media. In *Proceedings of the International Workshop on Fuzzy Logic and Applications*. 197–207.

Marco Viviani and Gabriella Pasi. 2017. Credibility in social media: Opinions, news, and health information—A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 7, 5 (2017), e1209.

Nguyen Vo and Kyumin Lee. 2018. The rise of guardians: Fact-checking URL recommendation to combat fake news. arXiv:1806.07516.

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

Amy B. Wang. 2016. 'Post-truth' named 2016 word of the year by Oxford Dictionaries. *Washington Post* (2016). Available at https://www.washingtonpost.com/news/the-fix/wp/2016/11/16/post-truth-named-2016-word-of-the-year-by-oxford-dictionaries/.

William Yang Wang. 2017. "Liar, liar pants on fire": A new benchmark dataset for fake news detection. arXiv:1705.00648.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, 849–857.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*.

Andrew Ward, L. Ross, E. Reed, E. Turiel, and T. Brown. 1997. Naive realism in everyday life: Implications for social conflict and misunderstanding. In *Values and Knowledge*, E. S. Reed, E. Turiel, and T. Brown (Eds.). The Jean Piaget Symposium Series. Lawrence Erlbaum Associates, 103–135.

Claire Wardle. 2017. Fake news. It's complicated. *First Draft News* 16 (2017). Available at https://firstdraftnews.org/latest/fake-news-complicated/.

Steven Euijong Whang and Hector Garcia-Molina. 2012. Joint entity resolution. In *Proceedings of the IEEE 28th International Conference on Data Engineering (ICDE'12)*. IEEE, Los Alamitos, CA, 294–305.

المنارة للاستشارات

www

Ke Wu, Song Yang, and Kenny Q. Zhu. 2015. False rumors detection on Sina Weibo by propagation structures. In *Proceedings of the IEEE 31st International Conference on Data Engineering (ICDE'15)*. IEEE, Los Alamitos, CA, 651–662.

Bishan Yang, Wen-Tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2014. Embedding entities and relations for learning and inference in knowledge bases. arXiv:1412.6575.

Fan Yang, Shiva K. Pentyala, Sina Mohseni, Mengnan Du, Hao Yuan, Rhema Linder, Eric D. Ragan, Shuiwang Ji, and Xia Ben Hu. 2019. XFake: Explainable fake news detector with visualizations. In *Proceedings of the World Wide Web Conference*. ACM, New York, NY, 3600–3604.

Junting Ye and Steven Skiena. 2019. MediaRank: Computational ranking of online news sources. arXiv:1903.07581.

Bowen Yu, Zhenyu Zhang, Tingwen Liu, Bin Wang, Sujian Li, and Quangang Li. 2019. Beyond word attention: Using segment attention in neural relation extraction. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 5401–5407. DOI: https://doi.org/10.24963/ijcai.2019/750

Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. 2014. *Social Media Mining: An Introduction*. Cambridge University Press.

Reza Zafarani, Xinyi Zhou, Kai Shu, and Huan Liu. 2019. Fake news research: Theories, detection strategies, and open problems. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, 3207–3208.

Dongsong Zhang, Lina Zhou, Juan Luo Kehoe, and Isil Yakut Kilic. 2016. What online reviewer behaviors really matter? Effects of verbal and nonverbal behaviors on detection of fake online reviews. *Journal of Management Information Systems* 33, 2 (2016), 456–481.

Jiawei Zhang, Limeng Cui, Yanjie Fu, and Fisher B. Gouza. 2018. Fake news detection with deep diffusive network model. arXiv:1805.08751.

Bin Zhou and Jian Pei. 2009. OSD: An online web spam detection system. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09)*, Vol. 9.

Denny Zhou, Olivier Bousquet, Thomas N. Lal, Jason Weston, and Bernhard Schölkopf. 2004a. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*. 321–328.

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

Xinyi Zhou, Atishay Jain, Vir V. Phoha, and Reza Zafarani. 2019a. Fake news early detection: A theory-driven model. arXiv:1904.11679.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. SAFE: Similarity-aware multi-modal fake news detection. In *Proceedings of the 24th Pacific-Asia Conference on Knowledge Discovery and Data Mining*.

Xinyi Zhou and Reza Zafarani. 2019. Network-based fake news detection: A pattern-driven approach. *ACM SIGKDD Explorations Newsletter* 21, 2 (2019), 48–60.

Xinyi Zhou, Reza Zafarani, Kai Shu, and Huan Liu. 2019b. Fake news: Fundamental theories, detection strategies and challenges. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining*. ACM, New York, NY, 836–837.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys* 51, 2 (2018), 1–36.

Miron Zuckerman, Bella M. DePaulo, and Robert Rosenthal. 1981. Verbal and nonverbal communication of deception. In *Advances in Experimental Social Psychology*. Vol. 14. Elsevier, 1–59.

Received December 2018; revised March 2020; accepted April 2020

المنارة للاستشارات

www

Copyright of ACM Computing Surveys is the property of Association for Computing Machinery and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.

www.manaraa.com