False News On Social Media: A Data-Driven Survey

Francesco Pierri
Politecnico di Milano
francesco.pierri@polimi.it

Stefano Ceri
Politecnico di Milano
stefano.ceri@polimi.it

ABSTRACT

In the past few years, the research community has dedicated growing interest to the issue of false news circulating on social networks. The widespread attention on detecting and characterizing false news has been motivated by considerable backlashes of this threat against the real world. As a matter of fact, social media platforms exhibit peculiar characteristics, with respect to traditional news outlets, which have been particularly favorable to the proliferation of deceptive information. They also present unique challenges for all kind of potential interventions on the subject.

As this issue becomes of global concern, it is also gaining more attention in academia. The aim of this survey is to offer a comprehensive study on the recent advances in terms of detection, characterization and mitigation of false news that propagate on social media, as well as the challenges and the open questions that await future research on the field. We use a data-driven approach, focusing on a classification of the features that are used in each study to characterize false information and on the datasets used for instructing classification methods. At the end of the survey, we highlight emerging approaches that look most promising for addressing false news.

1 INTRODUCTION

This section serves as an introduction to the topic of false news on social media; we provide some terminology, describe the social media platforms where false news are most widespread, overview psychological and social factors that are involved, discuss some of the effects on the real world and some open challenges. Finally, we discuss the focus of our survey in comparison with other existing surveys.

1.1 Terminology

In recent years, the terms **false news** and **fake news** have been broadly and interchangeably used to indicate information which can take a variety of flavors: disinformation, misinformation, hoaxes, propaganda, satire, rumors, clickbait and junk news. We provide next a list of the definitions encountered in the literature, which is by no means exhaustive. While there is common agreement that these terms indicate deceptive information, we believe that an agreed and precise definition is still missing.

Some researchers define **false news** as news articles that are potentially or intentionally misleading for the readers, as they are verifiable and deliberately false [3, 56]. They can represent fabricated information which mimics traditional news content in form but not in the intent or the organizational process [27]. It has been highlighted how the neologism **fake news** is usually employed with a political connotation with respect to the more traditional *false news* [27, 68].

Misinformation is defined as information that is inaccurate or misleading [27]. It could spread unintentionally [13] due to honest reporting mistakes or incorrect interpretations [12, 18]. In contrast, **disinformation** is false information that is spread deliberately to deceive people [27] or promote biased agenda [66].

Similarly to disinformation, **hoaxes** are intentionally conceived to deceive readers; qualitatively, they are described as *humorous and mischievous* (as defined in The Oxford English Dictionary) [26].

Satirical news are written with the primary purpose of entertaining or criticize the readers, but similarly to hoaxes they can be harmful when shared out of context [7, 47]. They are characterized by humor, irony and absurdity and they can mimic genuine news [48].

Propaganda is defined as information that tries to influence the emotions, the opinions and the actions of target audiences by means of deceptive, selectively omitting and one-sided messages. The purpose can be political, ideological or religious [65, 66].

Click-bait is defined as low quality journalism which is intended to attract traffic and monetize via advertising revenue [66].

The term **junk news** is more generic and it aggregates several types of information, from propaganda to hyperpartisan or conspiratorial news and information. It usually refers to the overall content that pertains to a publisher rather than a single article [71].

Finally, we came across several different definitions for **rumor**. Briefly, a rumor can be defined as a claim which did not originate from news events and that has not been verified while it spreads from one person to another [3, 56, 59]. As there exists a huge literature on the subject, we refer the interested reader to [74] for an extensive review.

1.2 Social media platforms as news outlets

The appearance of false news on news outlets is by no means a new phenomenon: in 1835 a series of articles published on the New York Sun, known as the *Great Moon Hoax*, described the discovery of life on the moon [3]. Nowadays the world is experiencing much more elaborated hoaxes; social media platforms have favored the proliferation of false news with much broader impact.

Most of nowadays **news consumption** has shifted towards online social media, where it is more comfortable to ingest, share and further discuss news with friends or other readers [17, 56, 58]. As producing content online is easier and faster, **barriers** for entering online media industry have dropped [3]. This has conveyed the dissemination of **low quality** news, which reject traditional journalistic standards and lack of third-party filtering and fact-checking [3]. These factors, together with a **decline** of general trust and confidence in traditional mass media, are the primary drivers for the explosive growth of false news on social media [3, 27].

Two main motivations have been proposed as to explain the rise of disinformation websites: 1) a **pecuniary** one, where viral news articles draw significant advertising revenue and 2) a more **ideological** one, as providers of false news usually aim to influence public opinion on particular topics [3]. Besides, the presence of **malicious agents** such as bots, cyborgs and trolls has been highlighted as another major cause to the spreading of misinformation [25, 53].

We refer the interested reader to [3] for an extensive analysis of all the factors explaining the spreading of false news in social media platforms.

1.3 Human factors

Aside from the technical aspects of social network platforms, the research community has leveraged a set of psychological, cognitive and social aspects which are considered as key contributors to the proliferation of false news on social media.

Humans have no natural expertise at distinguishing real from false news [26, 56]. Two major psychological theories explain this difficulty, respectively called **naive realism** and the **confirmation bias**. The former refers to the tendency of users to believe that their view is the only accurate one, whereas those who disagree are biased or uninformed [44]. The latter, also called *selective exposure*), is the inclination to prefer (and receive) information which confirms existing view [34]. As a consequence, presenting factual information to correct false beliefs is usually unhelpful and may increase misperception [35].

Some studies also mention the importance of **social identity theory** [5] and **normative social influence** [4]; accordingly, users tend to perform actions which are socially safer, thus consuming and spreading information items that agree with the norms established within the community.

All these factors are related to a certain extent to the well-known **echo chamber** (or *filter bubble*) effect, which gives rise to the formation of homogeneous clusters where individuals are similar people, that share and discuss similar ideas. These groups are usually characterized by extremely polarized opinions as they are insulated from opposite views and contrary perspectives [37, 59, 60]; it has been shown that these close-knit communities are the primary driver of misinformation diffusion [9].

Social technologies amplify these phenomenon as result of **algorithmic bias**, as they promote personalized content based on the preferences of users with the unique goal of maximize engagement [13, 27].

1.4 Effects on the real world

We can explain the explosive growth of attention on false news in light of a series of striking effects that the world has recently experienced.

Politics indeed accounts for most of the attention on false news. The 2016 US presidential elections have officially popularized the term *fake news* to the degree that it has been suggested that Donald Trump may not have been elected president were it not for the effects of false news (and the alleged interference of Russian trolls) [3]. Likewise, recent studies have shown that false news have also impacted on 2016 UK Brexit referendum [22] and the 2017 France presidential elections [14].

Over and above we may recall the **finance** stock crisis caused by a false tweet concerning president Obama [43], the **shootout** occurred in a restaurant as a consequence of the Pizzagate fake news [56] and the diffused mistrust towards **vaccines** during Ebola and Zika epidemics [15, 31].

1.5 Challenges

We mention here a few challenges which characterize the fight against false news on social media, as highlighted by recent research on the subject.

Firstly, false news are deliberately created to deceive the readers and to mimic traditional news outlets, resulting in an **adversarial scenario** where it is very hard to distinguish true news articles from false ones [53, 56].

Secondly, the **rate** and the **volumes** at which false news are produced overturn the possibility to fact-check and verify all items in a rigorous way, i.e. by sending articles to human experts for verification [53]. This also raises concern on developing tools for the **early detection** of false news as to prevent them from spreading in the network [28].

Finally, social media platforms impose limitations [54] on the **collection** of public data and as of today the community

has produced very limited training datasets, which typically do not include all the information relative to false news.

1.6 Survey Focus

Aside from a few seminal works appeared in 2015 and 2016 [7, 47, 48], we build our survey with a focus on the last two years, as most of the research on false news has developed in 2017 and 2018. Moreover, we concentrate on a few social networks which attracted most of the research focus: **Twitter**, **Facebook** and **Sina Weibo**[1]. This is mainly due to the public availability of data and the existence of proprietary application programming interfaces (API) which ease the burden of collecting data.

As our analysis is focused on the aforementioned social media, issues concerning false news on **collaborative platforms** such as Wikipedia and Yelp (namely fake reviews, spam detection, etc.) are out of the scope of this survey; we thus refer the reader to [25] for an overview of related research. We suggest [29] for a comprehensive review of the research that focuses, instead, on **rumors detection and resolution**, as we observed that many aspects are shared with our subject. Finally, we suggest [15] to the readers who may be interested in the research on **social bots**.

2 PROBLEM FORMULATION AND METHODOLOGY

Our presentation of research about false news on social media is divided into three parts. We first describe a huge body of works whose objective is to detect false news, then we describe works that explain the models of diffusion of false news and finally works that attempt to mitigate their effects.

We start our survey by considering a variegated landscape of research contributions which focus on the **detection** of false news. Their taxonomy, presented in Table 1, is based on three aspects: employed technique, considered features, and dataset used (referring to Appendix A).

The problem has been traditionally formulated as a supervised binary classification problem, starting with datasets consisting of labeled news articles, related tweets and Facebook posts which allow to capture different features, from content based ones (text, image, video) to those pertaining to the social context (diffusion networks, users' profile, metadata) and, in some cases, to external knowledge bases (Wikipedia, Google News). Labels carrying the classification into true and false news are typically obtained via fact-checking organizations or by manual verification of researchers themselves. Appendix A comparatively describes the datasets used as ground truth for false news classification.

[1]A popular Chinese microblogging website which is a hybrid between Facebook and Twitter.

For what concerns the classification method, a wide range of techniques are used, from traditional machine learning (Logistic Regression, Support Vector Machines, Random Forest) to deep learning (Convolutional and Recurrent Neural Networks) and to other models (Matrix Factorization, Bayesian Inference).

Despite the vast amount of contributions which belong to this area, we believe that false news detection requires a deeper and more structured approach. Several works appear as academic exercises, not always compared to each other (and often not comparable). They do not clearly express the impact of their results or the possible consequences in terms of real world effects.

Section 4 describes the literature which focuses on the **characterization** of misinformation spreading on social media. This is achieved by reconstructing the diffusion networks pertaining to false news, as resulting from multiple users interactions on the platforms. We describe these works in detail, as we believe that they provide substantial research contributions, also due to the wide reach of large-scale experiments carried out in these studies.

Finally, Section 5 presents a few works which tackle the problem of **mitigation** against false news on social media. Since main social networking platforms, from Facebook to Twitter, have recently provided to their users tools to combat disinformation [23], the general approach of research contributions in the field is to resort to the *wisdom of the crowd* as to identify malicious news items which, once verified by fact-checkers, may prevent from misinformation spreading on the platform.

3 FALSE NEWS DETECTION

We approach these methods by starting from those contributions which focus only on content-based features; we next describe contributions which consider only the social context and finally those that consider both aspects.

3.1 Content-based

Wang et al. (2017) [69] first introduced the Liar dataset (cfr. A.8) including textual and metadata features (such as the speaker affiliation or the source newspaper). They solved a multi-classification problem based on the six degrees of truth provided by the PolitiFact[2] fact-checking organization, using several machine learning and deep learning methods, from logistic regression to convolutional and recurrent neural networks.

A deep textual analysis is carried out in *Horne at al. (2017)* [20], where authors examine the body and title (cfr. A.1) of different categories of news articles (true, false and satire), extracting complexity, psychological and stylistic features.

[2]https://www.politifact.com/

		References															
		[69]	[20]	[41]	[40]	[42]	[11]	[21]	[61]	[66]	[70]	[72]	[28]	[73]	[49]	[57]	[65]
Techniques	Machine Learning	✓	✓		✓	✓	✓		✓							✓	✓
	Deep Learning	✓		✓						✓	✓	✓	✓		✓		✓
	Others						✓	✓	✓			✓		✓		✓	
Features	News Content	✓	✓	✓	✓	✓	✓	✓							✓	✓	✓
	Social Context								✓	✓	✓	✓	✓	✓	✓	✓	✓
	Dataset	A.8	A.1	A.3, A.9	A.4	A.2		A.7					A.10	A.8, A.1	A.10	A.5	

Table 1: Comparative description of seventeen studies for false news detection, in terms of method, considered features, and dataset

They highlight the relevance of each set in distinct classification tasks, using a linear Support Vector Machine (SVM), finally inferring that real news are substantially different from false news in title whereas satire and false news are similar in content. They also hypothesize the higher persuasion of false news in terms of an Elaboration Likelihood Model [36], observing that consuming false news requires little energy and cognition, making them more appealing to the readers.

A neural network model is also presented by *Popat et al. (2018)* [41], who build a framework to detect false claims and provide evidence for the credibility assessment via comprehensible snippets. They evaluate their model against some state-of-the-art techniques on different collections of news articles (cfr. A.3 and A.9) and they show examples of explainable results enabled by the *attention mechanism* embedded in the model, which highlights the words in the text that are more relevant for the classification outcome.

Perez-Rosas et al. (2018) [40] conduct a series of experiments to build different false news detectors on top of several sets of linguistic features (extracted from the body of news articles) namely ngrams, LIWC [39], punctuation, syntax and readability. They produce a dataset of false and true news articles (cfr. A.4) where they evaluate a linear SVM classifier, showing different performances depending on the features considered. They suggest that computational linguistics can effectively aide in the process of automatic detection of false news.

The goal of *Potthast et al. (2018)* [42] is to assess the style similarity of several categories of news, notably hyper-partisan, mainstream, satire and false. The methodology proposed employs an algorithm called Unmasking [24], which is a meta learning approach originally intended for authorship verification. They carry out several experiments comparing topic and style-based features with a Random Forest classifier and they conclude that, while hyper-partisan, satire and mainstream are well distinguished, a style-based analysis alone is not effective for detecting false news.

Fairbanks et al. (2018) [11] claim that modeling only text content of articles is sufficient to detect bias but it is not enough to identify false news. To this extent they compare two different models, a content-based one which uses a classifier on traditional textual features and a structural method that applies loopy belief propagation [33] on a graph which is built on the link structure of news articles. The dataset employed is a collection of articles which are gathered from the Global Database of Events Language and Tone (GDELT[3]) whereas labels are manually crawled from a fact-checking website.

Hosseini et al. (2018) [21] tackle the problem of distinguishing different categories of false news (from satire to junk news), based only on the content. Their approach involves a tensor decomposition of documents which aims to capture latent relationships between articles and terms and the spatial/contextual relations between terms. They further use an ensemble method to leverage multiple decompositions in order to discover classes with higher homogeneity and lower outlier diversity. They employ the Kaggle dataset (cfr. A.7) where they discriminate up to six different categories and they compare their algorithm with other state-of-the-art clustering techniques.

3.2 Context-based

Tacchini et al. (2017) [61] propose a technique to identify false news on the basis of users who *liked* them on Facebook. They collect a set of posts and users from both conspiracy theories and scientific pages and they build a dataset where each feature vector represents the set of users who *liked* a page. They eventually compare logistic regression with a (boolean crowdsourcing) harmonic algorithm for showing that they are able to achieve high accuracy with a little percentage of the entire training data.

Volkova et al. (2017) [66] address the problem of predicting four sub-types of suspicious news: satire, hoaxes, click-bait and propaganda. They start from a (manually constructed) list of trusted and suspicious Twitter news accounts and they collect a set of tweets in the period of Brussels bombing in 2016. They incorporate tweet text, several linguistic cues (bias, subjectivity, moral foundations) and user interactions in a *fused* neural network model which is compared against

[3]https://www.gdeltproject.org/

ad-hoc baselines trained on the same features. They qualitatively analyze the characteristics of different categories of news observing the performances of the model.

Wang et al. (2018) [70] propose a multi-modal neural network model which extracts both textual and visual features from tweets and Weibo conversations in order to detect false news items. Inspired by adversarial settings [16] they couple it with an *event discriminator*, which they claim is able to remove event-specific features and generalize to unseen scenarios, where the number of events is specified as a parameter. They evaluate the model on two custom datasets, but they compare it with ad-hoc baselines which are not conceived for false news detection.

Wu et al. (2018) [72] instead concentrate on modelling the propagation of messages carrying malicious items in social networks. They first infer embeddings for users from the social graph and in turn use a neural network model to classify news items. To this extent they provide a new model to embed a social network graph in a low-dimensional space and they construct a sequence classifier using *Long Short-Term Memory* (LSTM) networks [19] in order to analyze propagation pathways of messages. Therefore they build a custom dataset, reflecting both true and false news, by leveraging the Twitter API and the fact-checking website Snopes[4] and they present a comparison of classification performances in terms of different state-of-the-art embedding methods.

Propagation of news items is also taken into account by *Yu et al. (2018)* [28], who use a combination of convolutional and *Gated Recurrent Units* (GRU) [6] to model diffusion pathways as multivariate time series, where each point corresponds to the characteristics of the user retweeting the news, and perform early detection of false news. The method is evaluated on two real world datasets of rumors (cfr. A.10) and compared against some state-of-the-art-techniques originally conceived for rumor resolution.

The first unsupervised approach is provided in *Yang et al. (2019a)* [73], where veracity of news and users' credibility are treated as latent random variables in a Bayesian network model, and the inference problem is solved by means of collapsed Gibbs sampling approach [46]. The method is evaluated on LIAR (cfr. A.8) and BuzzFeedNews (cfr. A.1) datasets and compared with general truth discovery algorithms not explicitly designed for false news detection.

3.3 Content and Context-based

The contribution of *Ruchansky et al. (2017)* [49] is a neural network model which incorporates the text of (false and true) news articles, the responses they receive in social networks and the source users that promote them. The model is tested on Twitter and Weibo rumors datasets (cfr. A.10) and it is evaluated against other techniques conceived for rumor detection. They finally present an analysis of users behaviours in terms of lag and activity showing that the source is a promising feature for the detection.

In *Shu et al. (2017)* [57] a tri-relationship among publishers, news items and users is employed in order to detect false news. Overall, user-news interactions and publisher-news relations are embedded using non-negative matrix factorization [38] and users credibility scores. Several different classifiers are built on top of the resulting features and performances are evaluated on the FakeNewsNet dataset (cfr. A.5) against other state-of-the-art information credibility algorithms. Results show that the social context could effectively exploited to improve false news detection.

Volkova et al. (2018) [65] focus on inferring different deceptive strategies (misleading, falsification) and different types of deceptive news (propaganda, disinformation, hoaxes). Extending their previous work [66], they collect summaries, news pages and social media content (from Twitter) that refer to confirmed cases of disinformation. Besides traditional content-based features (syntax and style) they employ psycho-linguistic signals, e.g. biased language markers, moral foundations and connotations, to train different classifiers (from Random Forests to neural networks) in a multi-classification setting. Final results show that falsification strategies are easier to identify than misleading and that disinformation is harder to predict than propaganda or hoaxes.

4 MODELS OF FALSE NEWS DIFFUSION

A first large-scale study on online misinformation is provided by *Del Vicario et al. (2016)* [9], who carry out a quantitative analysis on news consumption relatively to scientific and conspiracy theories news outlets on Facebook. They leverage Facebook Graph API in order to collect a 5-years time span of all the posts (and user interactions) which belong to the aforementioned categories. They analyze cascades (or *sharing trees*) in terms of lifetime, size and edge homogeneity (i.e. an indicator of the polarization of users involved) and they show that 1) the consumption patterns differ in the two categories and that 2) the *echo chambers* (or communities of interest) appear as the preferential drivers for the diffusion of content. On top of these results, they build a data driven percolation model which accounts for homogeneity and polarization and they simulate it in a small-world network reproducing the observed dynamics with high accuracy.

Similarly, a groundbreaking contribution is provided in *Vosoughi et al. (2018)* [68], where the entire Twitter universe is explored in order to track the diffusion of false and true news. Authors build a collection of links to fact-checking articles (from six different organizations) which correspond to true, false and mixed news stories and they accordingly

[4] https://www.snopes.com/

investigate how these rumors spread on the Twitter network by gathering only tweets that explicitly contain the URLs of the articles. The resulting dataset contains approx. 126000 stories tweeted by 3 millions users more than 4.5 million times. A series of measures are carried out including statistical and structural indicators of the retweeting networks along with sentiment analysis, topic distribution and novelty estimation of the different categories of news. The final results show that overall falsehood spread significantly faster, deeper, farther and broader than the truth in all categories of information, with a prominent weight on political news. Moreover, they observe that false news usually convey a higher degree of novelty and that novel information is more likely to be shared by users (although they cannot claim this is the only reason behind the "success" of misinformation).

A slightly diverse analysis is issued in *Shao et al. (2018a)* [54], where authors study the structural and dynamic characteristics of the core of the diffusion network on Twitter before and after the 2016 US Presidential Elections. They first illustrate the implementation and deployment of the Hoaxy platform [52] which is then employed to gather the data required for their analysis. They build different datasets (relative to a few months before and after the elections) which correspond to fact-checking and misinformation articles, i.e. the retweeting network of users that share URLs for related news items, and they perform a k-core decomposition analysis to investigate the role of both narratives in the network. They show that low-credibility articles prevail in the core, whereas fact-checking is almost relegated to the periphery of the network. They also carry out a network robustness analysis in order to analyze the role of most central nodes and guide possible different interventions of social platforms.

Same authors largely extend previous results in *Shao et al. (2018b)* [53], as they carry out a huge analysis on Twitter in a period of ten months in 2016 and 2017. They aim to find evidence of the considerable role of social bots in spreading low-credibility news articles. The Hoaxy [52] platform is leveraged once again and more than 14 millions tweets, including fact-checking and misinformation sources, are collected. The *Botometer* algorithm [8] is used to assess the presence of social bots among Twitter users. The results show that bots are active especially in the first phase of the diffusion, i.e. a few seconds after articles are published, and that although the majority of false articles goes unnoticed, a significant fraction tends to become viral. They also corroborate, to a certain extent, results provided by *Vosoughi et al. (2018)* [68]. Moving on, they highlight bot strategies for amplifying the impact of false news and they analyze the structural role of social bots in the network by means of a network dismantling procedure [2]. They finally conclude that most influential nodes are likely to be bots and that curbing such accounts would yield the second most effective strategy to reduce misinformation. Alternatively CAPTCHAs [67] are indicated as a simple tool to distinguish bots from humans but they would also add undesirable effects to the user experience of the platform.

Differently from previous works, a study of the agenda-setting [30] power of false news is instead accomplished in *Vargo et al. (2018)* [63], where authors focus on the online mediascape from 2014 to 2016. They leverage a few different agenda-setting models with a computational approach (collecting data from GDELT) in order to examine, among other targets, the influence of false news on real news reports, i.e. whether and to which extent false news have shifted journalistic attention in mainstream, partisan and fact-checking organizations. To this extent they gather news articles corresponding to partisan and mainstream news outlets as well as fact-checking organizations and false news websites; they refer to diverse references in the literature in order to manually construct the list. A network of different events and themes (as identified in the GDELT database) is built to relate distinct media and to model time series of (eigenvector) centrality scores [50] in order to carry out Granger causality tests and highlight potential correlations. Besides other results, they show that partisan media indeed appeared to be susceptible to the agendas of false news (probably because of the elections) and that the agenda setting power of false news, which has more freedom than ever, is declining.

5 FALSE NEWS MITIGATION

Finally, a few potential interventions have been proposed for reducing the spread of misinformation on social platforms, from curbing most active (and likely to be bots) users [53] to leveraging the users' flagging activity in coordination with fact-checking organizations. The latter approach is proposed as a first practical mitigation technique in [23] and [62], where the goal is to reduce the spread of misinformation leveraging users' flagging activity on Facebook.

Kim et al. (2018) [23] develop CURB, an algorithm to select the most effective stories to send for fact-checking as to efficiently reduce the spreading of non-credible news with theoretical guarantees; they formulate the problem in the context of temporal point processes [1] and stochastic differential equations and they use the Rumors datasets (A.10) to evaluate it in terms of precision and misinformation reduction (i.e. the fraction of prevented unverified exposures). They show that the algorithm accuracy is very sensitive to the ability of the crowd at spotting misinformation.

Tschiatschek et al. (2018) [62] also aim to select a small subset of news to send for verification and prevent misinformation from spreading; however, as they remark, with a few differences from the previous method respectively 1) they learn the accuracy of individual users rather than considering all of them equally reliable and 2) they develop an

algorithm which is agnostic to the actual propagation of news in the network. Moreover, they carry out their experiments in a simulated Facebook environment where false and true news are generated by users in a probabilistic manner. They show that they are able at once to learn users' flagging behaviour and consider possible adversarial behaviour of spammer users who want to promote false news.

A different contribution is issued by *Vo et al. (2018)* [64], who are the first to examine active Twitter users who share fact-checking information in order to correct false news in online discussions. They incidentally propose a URL recommendation model to encourage these *guardians* (users) to engage in the spreading of credible information as to reduce the negative effects of misinformation. They use Hoaxy [52] (cfr. A.6) to collect a large amount of tweets referring to fact-checking organizations and they analyze several characteristics of the users involved (activity, profile, topics discussed, etc). Finally they compare their recommendation model, which takes into account the social structure, against state-of-the-art collaborative filtering algorithms.

6 DISCUSSION AND CONCLUSIONS

Despite the vast review of literature presented so far, in agreement with [27] we believe that there are only a few substantial research contributions, most of which specifically focus on characterizing the diffusion of misinformation on social media.

Although different psycho-linguistic signals derived from textual features are useful for false news detection, content alone may not be sufficient and other features, inferred from the social dimension, should be taken into account in order to distinguish false news from true news.

The lack of gold-standard agreed datasets and of research guidelines on the subject has favored the diffusion of ad-hoc data collections; the related detection techniques share several limitations, as they do not always compare with each other and do not explicitly discuss the impact and consequences of their results.

Besides the existing challenges highlighted in the introductory section, we believe that: 1) in light of recent contributions on false news diffusion networks, more insights for false news detection should be gained from network analysis; 2) future research should overcome the limitations imposed by manual fact-checking and address the problem in a unsupervised manner; 3) in general, the research community should coordinate efforts originating from different areas (from psychology to journalism to computer science) in a more structured fashion; 4) future contributions should favour the development of real world applications for providing effective help in the fight against false news.

REFERENCES

[1] O. Aalen, O. Borgan, and H. Gjessing. *Survival and event history analysis: a process point of view*. Springer Science & Business Media, 2008.

[2] R. Albert, H. Jeong, and A.-L. Barabási. Error and attack tolerance of complex networks. *Nature*, 406(6794):378, 2000.

[3] H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36, 2017.

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

[5] B. E. Ashforth and F. Mael. Social identity theory and the organization. *Academy of management review*, 14(1):20–39, 1989.

[6] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[7] N. J. Conroy, V. L. Rubin, and Y. Chen. Automatic deception detection: Methods for finding fake news. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, page 82. American Society for Information Science, 2015.

[8] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 273–274. International World Wide Web Conferences Steering Committee, 2016.

[9] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559, 2016.

[10] L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G. W. S. Hoi, and A. Zubiaga. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, 2017.

[11] J. Fairbanks et al. Credibility assessment in the news: Do we need to read? 2018.

[12] D. Fallis. A conceptual analysis of disinformation. *iConference*, 2009.

[13] M. Fernandez and H. Alani. Online misinformation: Challenges and future directions. In *Companion of the The Web Conference 2018 on The Web Conference 2018*, pages 595–602. International World Wide Web Conferences Steering Committee, 2018.

[14] E. Ferrara. Disinformation and social bot operations in the run up to the 2017 french presidential election. *First Monday*, 22(8), 2017.

[15] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini. The rise of social bots. *Communications of the ACM*, 59(7):96–104, 2016.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[17] J. Gottfried and E. Shearer. *News Use Across Social Medial Platforms 2016*. Pew Research Center, 2016.

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

[19] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[20] B. D. Horne and S. Adali. This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *arXiv preprint arXiv:1703.09398*, 2017.

[21] S. Hosseinimotlagh and E. E. Papalexakis. Unsupervised content-based identification of fake news articles with tensor decomposition ensembles. 2018.

[22] P. N. Howard and B. Kollanyi. Bots,# strongerin, and# brexit: computational propaganda during the uk-eu referendum. *arXiv preprint arXiv:1606.06356*, 2016.

[23] J. Kim, B. Tabibian, A. Oh, B. Schölkopf, and M. Gomez-Rodriguez. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 324–332. ACM, 2018.

[24] M. Koppel, J. Schler, and E. Bonchek-Dokow. Measuring differentiability: Unmasking pseudonymous authors. *Journal of Machine Learning Research*, 8(Jun):1261–1276, 2007.

[25] S. Kumar and N. Shah. False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559, To appear in the book titled Social Media Analytics: Advances and Applications, by CRC press, 2018*, 2018.

[26] S. Kumar, R. West, and J. Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*, pages 591–602. International World Wide Web Conferences Steering Committee, 2016.

[27] D. M. J. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts, and J. L. Zittrain. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[28] Y. Liu and Y.-F. Wu. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. *AAAI Conference on Artificial Intelligence*, 2018.

[29] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 3818–3824. AAAI Press, 2016.

[30] M. McCombs. *Setting the agenda: Mass media and public opinion*. John Wiley & Sons, 2018.

[31] J. Millman. The inevitable rise of ebola conspiracy theories. *The Washington Post*, 2014.

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

[33] K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 467–475. Morgan Kaufmann Publishers Inc., 1999.

[34] R. S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175, 1998.

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

[36] D. J. O'Keefe. Elaboration likelihood model. *The international encyclopedia of communication*, 2008.

[37] E. Pariser. *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011.

[38] V. P. Pauca, F. Shahnaz, M. W. Berry, and R. J. Plemmons. Text mining using non-negative matrix factorizations. In *Proceedings of the 2004 SIAM International Conference on Data Mining*, pages 452–456. SIAM, 2004.

[39] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn. The development and psychometric properties of LIWC2015. Technical report, 2015.

[40] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401. Association for Computational Linguistics, 2018.

[41] K. Popat, S. Mukherjee, A. Yates, and G. Weikum. Declare: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, 2018.

[42] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240. Association for Computational Linguistics, 2018.

[43] K. Rapoza. Can 'fake news' impact the stock market? *Forbes*, 2017.

[44] E. S. Reed, E. Turiel, and T. Brown. Naive realism in everyday life: Implications for social conflict and misunderstanding. In *Values and knowledge*, pages 113–146. Psychology Press, 2013.

[45] M. Risdal. Fake news dataset. https://www.kaggle.com/mrisdal/fake-news. 2017.

[46] C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

[47] V. Rubin, N. Conroy, Y. Chen, and S. Cornwell. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17, 2016.

[48] V. L. Rubin, Y. Chen, and N. J. Conroy. 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*, page 83. American Society for Information Science, 2015.

[49] N. Ruchansky, S. Seo, and Y. Liu. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 797–806. ACM, 2017.

[50] B. Ruhnau. Eigenvector centrality: a node centrality? *Social networks*, 22(4):357–365, 2000.

[51] G. Santia and J. Williams. Buzzface: A news veracity dataset with facebook user commentary and egos. *International AAAI Conference on Web and Social Media*, 2018.

[52] C. Shao, G. L. Ciampaglia, A. Flammini, and F. Menczer. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, pages 745–750, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

[53] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer. The spread of low-credibility content by social bots. *Nature communications*, 9(1):4787, 2018.

[54] C. Shao, P.-M. Hui, L. Wang, X. Jiang, A. Flammini, F. Menczer, and G. L. Ciampaglia. Anatomy of an online misinformation network. *PLOS ONE*, 13(4):1–23, 04 2018.

[55] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*, 2018.

[56] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36, Sept. 2017.

[57] K. Shu, S. Wang, and H. Liu. Beyond news contents: The role of social context for fake news detection. *arXiv preprint arXiv:1712.07709 (2017), to appear in Proceedings of 12th ACM International Conference on Web Search and Data Mining (WSDM 2019)*.

[58] C. Silverman. This analysis shows how fake election news stories outperformed real news on facebook. BuzzFeed, https://zenodo.org/record/1239675, 2016.

[59] C. Sunstein. On rumors: How falsehoods spread, why we believe them, what can be done., 2007.

[60] C. R. Sunstein. *Echo chambers: Bush v. Gore, impeachment, and beyond*. Princeton University Press, 2001.
[61] E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret, and L. de Alfaro. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506*, 2017.
[62] S. Tschiatschek, A. Singla, M. Gomez Rodriguez, A. Merchant, and A. Krause. Fake news detection in social networks via crowd signals. In *Companion of the The Web Conference 2018 on The Web Conference 2018*, pages 517–524. International World Wide Web Conferences Steering Committee, 2018.
[63] C. J. Vargo, L. Guo, and M. A. Amazeen. The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *New Media & Society*, 20(5):2028–2049, 2018.
[64] N. Vo and K. Lee. The rise of guardians: Fact-checking url recommendation to combat fake news. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 275–284, New York, NY, USA, 2018. ACM.
[65] S. Volkova and J. Y. Jang. Misleading or falsification: Inferring deceptive strategies and types in online news and social media. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 575–583, 2018.
[66] S. Volkova, K. Shaffer, J. Y. Jang, and N. Hodas. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 647–653, 2017.
[67] L. Von Ahn, M. Blum, N. J. Hopper, and J. Langford. Captcha: Using hard ai problems for security. In *International Conference on the Theory and Applications of Cryptographic Techniques*, pages 294–311. Springer, 2003.
[68] S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.
[69] W. Y. Wang. " liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 422–426, 2017.
[70] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 849–857. ACM, 2018.
[71] S. C. Woolley and P. N. Howard. *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media*. Oxford University Press, 2018.
[72] L. Wu and H. Liu. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 637–645. ACM, 2018.
[73] S. Yang, K. Shu, S. Wang, R. Gu, F. Wu, and H. Liu. Unsupervised fake news detection on social media: A generative approach. In *Proceedings of 33rd AAAI Conference on Artificial Intelligence*, 2019.
[74] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter. Detection and resolution of rumours in social media: A survey. *ACM Comput. Surv.*, 51(2):32:1–32:36, Feb. 2018.

A DATASETS

The research community has produced a rich but heterogeneous ensemble of data collections for fact checking, often conceived for similar objectives and for slightly different tasks. We first introduce the datasets which are referenced in this survey along with a short description, the source and the main references; their features are summarized in Table 2. Next, we present some other interesting datasets.

A.1 BuzzFeedNews

BuzzFeed[5] News journalists have produced different collections of verified false and true news, shared by both hyperpartisan and mainstream news media on Facebook in 2016 and 2017; two of them, introduced by *Silverman (2016)* [58], consist of title and source of news items and they are used in [20, 51, 73]

A.2 BuzzFeed-Webis

This collection extends the previous one as it also contains the full content of shared articles with attached multimedia; it is employed in [42].

A.3 DeClare

This dataset contains several articles from Snopes, PolitiFact and NewsTrust [32] corresponding to both true and false claims; it is proposed in [41] and used for false news detection.

A.4 FakeNewsAMT

This collection contains some legitimate articles from mainstream news, some false news generated by Amazon Mechanical Turk workers and some false and true claims from GossipCop[6] (a celebrity fact-checking website); it is introduced in [40] for false news detection.

A.5 FakeNewsNet

This dataset contains both news content (source, body, multimedia) and social context information (user profile, followers/followee) regarding false and true articles, collected from Snopes and BuzzFeed and shared on Twitter; it was presented in [55] and employed in [57].

A.6 Hoaxy

The Hoaxy platform[7] has been first introduced in [52] and employed in several studies [53, 54, 64] for different goals; it is continuously monitoring the diffusion network (on Twitter, since 2016) of news articles from both disinformation and fact-checking websites and it allows to generate custom data collections.

A.7 Kaggle

This dataset was conceived for a Kaggle false news detection competition [45] which contains text and metadata from

[5] https://www.buzzfeed.com
[6] https://www.gossipcop.com
[7] https://hoaxy.iuni.iu.edu

	Content Features	Social Context Features	Size	Labeling	Platform	Reference
BuzzFeedNews	Article title and source	Engagement ratings	10^2	BuzzFeed	Facebook	[58]
BuzzFeedWebis	Full Article	-	10^3	BuzzFeed	Facebook	[42]
DeClare	Fact-checking post	-	10^5	NewsTrust PolitiFact Snopes	-	[41]
FakeNewsAMT	Article text only	-	10^3	Manual GossipCop	-	[40]
FakeNewsNet	Full article	Users metadata	10^3	BuzzFeed PolitiFact	Twitter	[55]
Hoaxy	Full article	Diffusion network Temporal trends Bot score (for users)	$> 10^6$	-	Twitter	[52]
Kaggle	Article text and metadata	-	10^4	BS Detector	-	[45]
Liar	Short statement	-	10^4	PolitiFact	-	[69]
SemEval-2017 Task8	Full article Wikipedia articles	Threads (tweets, replies)	10^4	Manual	Twitter	[10]
Rumors	Fact-checking title	Diffusion network (Twitter) Original message, replies (Weibo)	10^4	Snopes Weibo	Twitter Sina Weibo	[29]

Table 2: Comparative description of the datasets referenced in this survey.

websites indicated in the BS Detector[8]; it is employed in [21].

A.8 Liar

This is a collection of short labeled statements from political contexts, collected from PolitiFact, which serve for false news classification; it first appeared in [69] and it is employed in [73].

A.9 SemEval-2017 Task8

This data collection, composed of tweets and replies which form specific *conversations*, was designed for the specific tasks of stance and veracity resolution of social media content on Twitter; it is described in [10] and used in [41].

A.10 Rumors

This dataset was originally conceived for rumor detection and resolution in Twitter and Sina Weibo; introduced in [29], it contains retweet and discussion cascades corresponding to rumors/non-rumors and it is employed for false news detection and mitigation in [23, 28, 49].

A.11 Others

BuzzFace is a novel data collection composed of annotated news stories that appeared on Facebook during September 2016; it extends previous BuzzFeed dataset(s) (cfr. A.1) with comments and the web-page content associated to each news article; it is introduced in [51].

As a complement to Hoaxy (cfr. A.6), **JunkNewsAggregator** is a platform that tracks the spread of disinformation on Facebook pages; it is developed by authors of [71].

Other datasets point to relevant organizations in the context of false news: [63] contains a list of false news outlets as indicated by different fact-checking organizations, whereas the list of signatories[9] of the International Fact Checking Network's code of principles is a collector of the main fact-checking organizations which operate in different countries. Finally, [3] provides a set of the most shared false articles identified on Facebook during 2016 US elections.

[8] https://github.com/bs-detector/bs-detector

[9] https://ifcncodeofprinciples.poynter.org/signatories