A Survey on Fake News and Rumour Detection Techniques

Alessandro Bondielli[a,b], Francesco Marcelloni[a]

[a] *Dipartimento di Ingegneria dell'Informazione, University of Pisa, Largo Lucio Lazzarino, 1, Pisa, Italy*
[b] *Dipartimento di Ingegneria dell'Informazione, University of Florence*

Abstract

False or unverified information spreads just like accurate information on the web, thus possibly going viral and influencing the public opinion and its decisions. Fake news and rumours represent the most popular forms of false and unverified information, respectively, and should be detected as soon as possible for avoiding their dramatic effects. The interest in effective detection techniques has been therefore growing very fast in the last years. In this paper we survey the different approaches to automatic detection of *fake news* and *rumours* proposed in the recent literature. In particular, we focus on five main aspects. First, we report and discuss the various definitions of fake news and rumours that have been considered in the literature. Second, we highlight how the collection of relevant data for performing fake news and rumours detection is problematic and we present the various approaches, which have been adopted to gather these data, as well as the publicly available datasets. Third, we describe the features that have been considered in fake news and rumour detection approaches. Fourth, we provide a comprehensive analysis on the various techniques used to perform rumour and fake news detection. Finally, we identify and discuss future directions.

Keywords:
Fake News, Rumours, Natural Language Processing, Data Mining, Text

Email addresses: `alessandro.bondielli@unifi.it` (Alessandro Bondielli), `francesco.marcelloni@unipi.it` (Francesco Marcelloni)

Mining, Classification, Machine Learning, Deep Learning

1. Introduction

Nowadays, internet has become an integral part of our lifestyle. The role of traditional information channels such as newspapers and television on how we collect and consume news has become less prominent than in the past. Certainly, the growth of social media platforms has played a crucial role in this transformation. In fact, social networks like Twitter[1] and Facebook[2] have registered an exponential spike in popularity. Facebook, for example, reported to have 2.07 billion monthly active users in November 2017. On average, 1.37 billion of these users employ Facebook daily [3]. Twitter had 330 millions people as of January 2018 [4]. These numbers have kept constantly growing since the launch of the respective platforms.

A lot of people use social media platforms not only to keep in touch with friends and family, but also to gather information and news from around the world. Thus, social media play a fundamental role in the news fruition. The case study for Britain reported in [70] shows an increase in the usage of social media, and more importantly their relevance to news consumption.

According to [111], social media have become a critical publishing tool for journalists [31, 90] and the main consumption method for citizens looking for the latest news [43]. Journalists may use social media to report on public opinions about breaking news stories, and even to discover potential new stories, whereas citizens may follow the development of breaking news and events through official channels (i.e. news outlets official accounts on social media platforms) or through posts of their own network (e.g. friends, family, public figures). Indeed, social networks have proved to be extremely useful especially during crisis sit-

[1] https://twitter.com

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

[3] https://newsroom.fb.com/company-info/

[4] https://www.omnicoreagency.com/twitter-statistics/

uations, because of their inherent ability to spread breaking news much faster than traditional media [93].

However, this positive impact of the social media comes at a cost: the absence of control and fact-checking over posts makes social media a fertile ground for the spread of unverified and/or false information [111]. People often publish posts or share other people's posts verifying neither the source nor the information validity and reliability. Oftentimes, an attractive headline is sufficient for an article to be shared thousands of times, despite a possibly unsubstantiated or false content. For example, [101] reports that in 2016 a fake news about Hillary Clinton leading a child sex trafficking ring led a man to assault with a rifle a pizzeria that was claimed as one of the hubs of such trafficking [52]. This is only one example, but countless fake news and rumours are spread through social media every day for different reasons.

Fake news are not, however, a product of the digital communication age. Before the advent of internet, journalists were tasked with verification and fact checking of news and sources [79], thus making the exposure of public opinion to fake news more limited. Nowadays, social media facilitate the spread of the unverified and false information among a larger number of users, thus deeply influencing the global perception and the understanding of events [111]. Probably, one of the most striking examples of how fake news can influence opinions has been the U.S. presidential campaign in 2016. Authors in [4] thoroughly studied the subject, reporting interesting findings: during the campaign voters were exposed to higher number of pro-Trump than pro-Clinton articles. However, it is unclear how fake news can have been effective in influencing the final vote. An analysis performed through surveys has however shown that Republican voters were in general more inclined to believe in both real and fake news articles [4]. Thus, in this particular case, this analysis suggests that the influence of fake news on the final vote was relatively low. Nonetheless, we can argue that fake news and, more broadly, disinformation are becoming a huge problem on the web, and might have an important social cost in the future.

For this reason, both social network platforms and the research community are very active in identifying potential fake claims and assessing their veracity. For example, Facebook published a small set of guidelines[5] that should be helpful in preventing users from falling victim of fake news posts.

Fake news can take several different forms and shapes in the social media environment, thus it is even more difficult to efficiently detect and contrast them, both manually and automatically. For example, *clickbait* headlines are often used to attract users to open possibly biased articles in order to gain money from views.

In the context of social media, another issue closely related to the fake news is the *rumours* problem. A rumour can be defined as an unverified claim, which is made by users on social media platforms and can potentially spread beyond their private network. The claim can turn out to be either true or false, with the last option being an obvious problem for the community. Similar to fake news, the spread of false rumours can cause severe damage even in a short period. In [63] an example is discussed where, based on a false rumour about a kidnapping of kids and shooting near schools in Veracruz, car crashes were caused by parents rushing to take their children.

In studying rumour spread on Twitter, the authors in [91] noted that, within a conversation on a topic, users express opinions on its veracity, potentially giving insight to other users. Experimental results presented in [116], however, have shown that true rumours tend to be resolved faster than fake ones, and that unverified rumours produce a distinctive burst in the number of re-tweets within the first few minutes considerably higher than the one generated by rumours proven to be true or false. Furthermore, users generally tend to support unverified rumours. This, combined with the structure of social media platforms, may enable a chain reaction that can lead to the spread of unverified and potentially fake information.

In the last few years, the interest of the research community on fake news and rumours is considerably grown. Figure 1 shows the number of papers concerning

[5] https://www.facebook.com/help/188118808357379

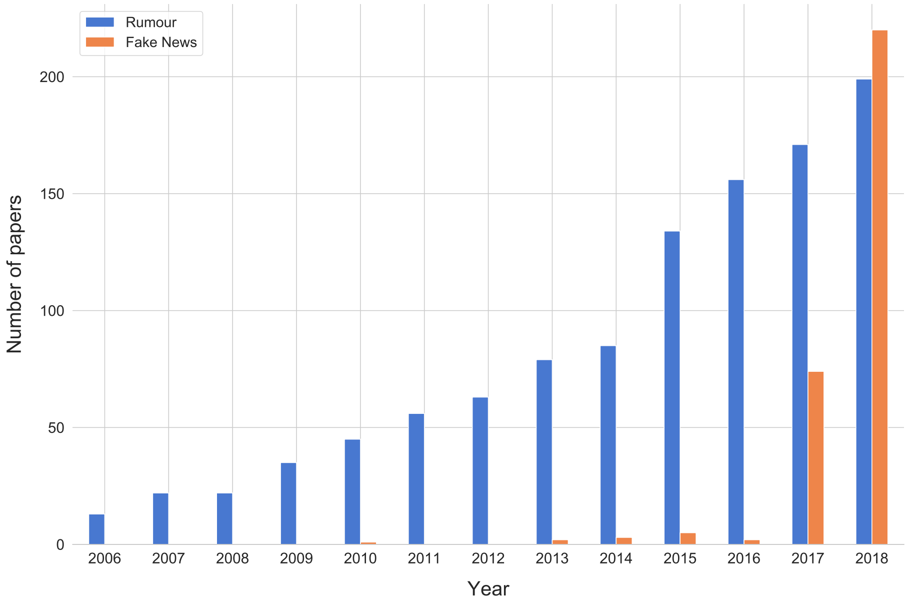

Figure 1: Trend of the number of papers published on Fake news and Rumours in the last years

rumors and *fake news* published from 2006 to 2018 and indexed in the Scopus database [6]. We can observe that the interest on rumours has been steadily growing over the past twelve years, while fake news has attracted an enormous interest mainly in the last two years, probably following the US presidential election in 2016 and the subsequent controversies on the role played by fake news on the win of the president Trump. The large number of papers published so far on the subject and the steadily increase of these papers in the last years motivate our work, which aims to identify current state of the art methods as well as critical issues and future trends. We can observe a similar trend also for papers that treat *fake news detection* and *rumour detection*, as shown in Figure 2, although starting only from 2011, when the phenomenon of the influence of fake news and rumours on the public opinion has started to be actually relevant.

[6] https://www.scopus.com/

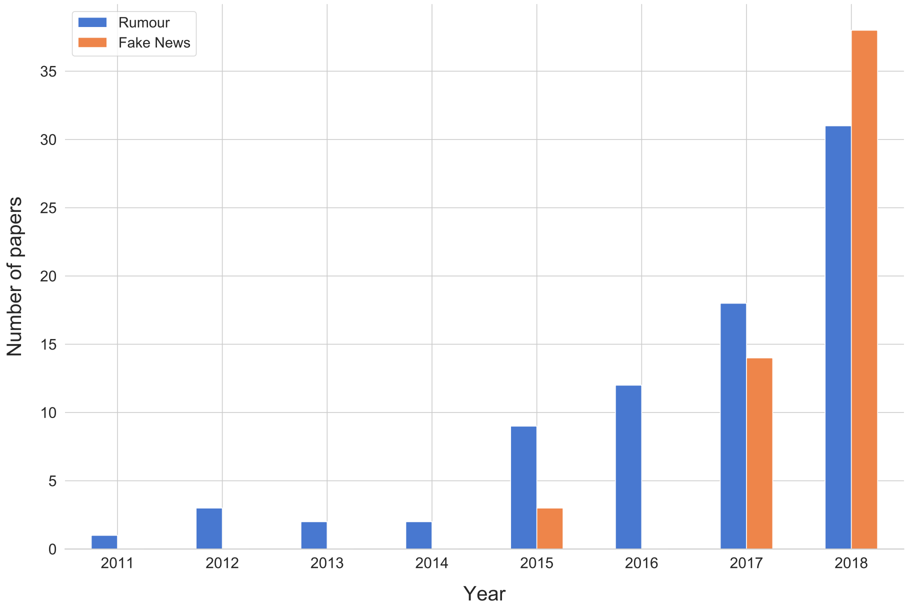

Figure 2: Trend of the number of papers published on Fake news detection and Rumour detection in the last 12 years

Many different aspects of false information on the web have been analysed by researchers with the aim of gaining insight on how this information spreads via social media channels and can be efficiently and rapidly detected, in order to reduce its impact on social network users and also on the society as a whole. Some of the most explored aspects are, for example, the possibility of inferring the veracity of a single piece of news (e.g. an article on a website), as well as identifying and debunking a false rumour that spreads through social media posts, based on information within it (i.e. text, authors, etc.) and surrounding it (i.e. network features, external knowledge, etc.).

Some reviews have been proposed in the last few years on fake news [85] and rumours [111] detection, dealing with the two issues separately. We believe, however, that such issues are strongly connected, and may be considered as different aspects of the same general problem. For this reason, the main contribution of the present survey is to provide a comprehensive analysis on simultaneously fake news and rumours detection under different perspectives. In particular, we focus on five main aspects: definitions, data collection, feature extraction and selection, techniques for analysis and detection, and future directions.

The paper is organized as follows. In Section 2 we provide an overview of various types of false information that can be found online, and introduce some definition. Section 3 discusses how data are collected and pre-processed for detecting fake news and rumours. In Section 4, we review the different types of features which have been extracted and used in the literature for detection. Section 5 introduces and discusses the different techniques used in recent years for fake news and rumour detection. In Section 6 we present ongoing problems and describe possible future research directions. Section 7 draws some conclusion.

2. False Information Basics

In this Section, we introduce some definition on key aspects of Internet-based false information as well as the description of different data sources of this information on the web. Obviously, the most common terms in mainstream media are *fake news* and *rumours*, but however researchers have also analyzed other aspects related to misinformation on the web, such as *clickbait*, *social spam* and *fake reviews*. In the literature, different categorizations of fake news and rumours have been proposed, mostly depending on source and type of data used for analysis. Early studies in this field, especially from a computational perspective, are relatively recent. Therefore, the boundaries of the study matter are often not clearly defined. For this reason, we believe that it is fundamental to provide some insight into what kind of data can become matter of analysis and how to define it.

Figure 3 shows a simple categorization of various types of misinformation. Although for the sake of completeness in the figure we report other types of misinformation, this paper focuses only on fake news and rumours.

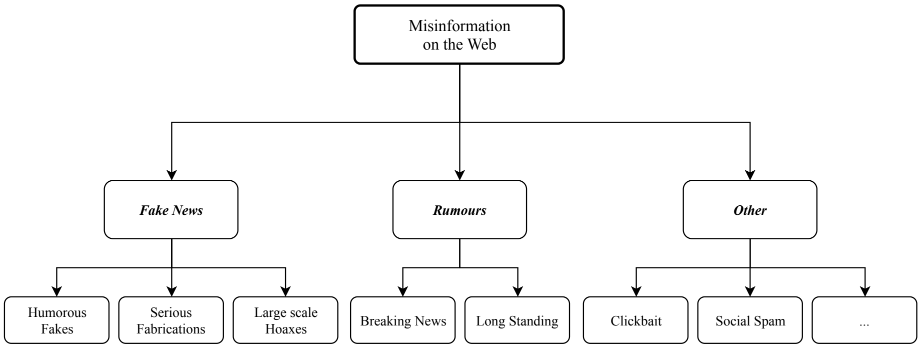

Figure 3: Categorization of various types of misinformation on the web.

2.1. Fake News

”Fake news” has become the de-facto expression for identifying false information in mainstream media, especially for web-related content, mostly spreading during and after the 2016 U.S. Presidential Campaign. However, research on fake news generally uses a more restrictive definition. Following [4], a fake news is “a news article that is intentionally and verifiably false”. Such definition hinges on two key aspects: *intent* and *verifiability*. Fake news are therefore news articles that are intentionally written to mislead or misinform readers, but can be verified as false by means of other sources. Several recent studies, such as [24], have adopted this definition.

In [79] a distinction among different aspects of fake news has been introduced. In particular, the authors focus on *serious fabrications*, *large scale hoaxes* and *humorous fakes*. *Serious fabrications* are the prototypical form of fake news, i.e. articles with a malicious intent (e.g. faked interviews, pseudoscience articles, etc.), that often become viral through social media. *Large scale hoaxes* are reports of false information disguised as proper news [79]. Usually such hoaxes are organized in a larger scale than a simple news article, often targeting public figures or ideas. Finally, *humorous fakes* are written in order to amuse readers, who are considered to be aware of the humorous intent of the author. Examples are satirical pieces masqueraded as real news, such as the ones produced by websites like *The Onion* and *lercio.it*

Concluding, we can identify three key aspects of fake news: i) its form, as news article; ii) its deceptive intent, that can be either satirical or malicious; and iii) the verifiability of its content as completely or partially false.

2.2. Rumours

In the recent scientific literature, *rumours* are probably the most widely studied false information on the web. They refer to information that has not been confirmed by official sources yet and is spread mostly by users on social media platforms.

Rumours are not a product of the Internet age, with the early studies dating back to the end of the World War II [5, 6]. However, we can argue that Internet and social media platforms in particular are a fertile ground for the spread of unconfirmed information [97].

To provide a formal definition of rumour is not straightforward. In fact, researchers have reported different interpretations. One of the most widely adopted definitions comes from the authors in [30]. In their research, rumours are identified as "unverified and instrumentally relevant information statements in circulation". Furthermore, [115] defines a rumour more specifically as a "circulating story of questionable veracity, which is apparently credible but hard to verify, and produces sufficient skepticism and/or anxiety". For this definition a rumour has to produce an impactful reaction on its audience. However, we can argue that these definitions hinge on the "unverified" characteristic of the information. This unverified information could be true, partly true, entirely false or remain unverified [111].

Other works, such as [14], have opted instead for defining rumours as circulating false information. In this case, the unverified aspect of information is disregarded, and the main focus is posed on its veracity. Methods and scope of such studies remain however similar to those adopting the former presented definition.

Finally, different studies have tried to classify rumours with respect to their type, scope, and characteristics. For example, [53] has categorized rumours with respect to the expected reaction in a psychological perspective. From a more practical standpoint, [111] has interestingly split rumours into two main categories. On the one hand, *long standing rumours* represent information that circulates for long periods of time and may never be verified as true or false. Urban legends and conspiracy theories can be considered as long standing rumour. For example, the rumour that Barack Obama is muslim has been studied in the literature. On the other hand, *breaking news rumours* are most common, and appear in connection with breaking news stories. They can be the product of unintentional misinformation, but could also prove to be deceptive in nature. Breaking news rumours have received more attention in the literature with respect to long standing rumours. This is due to the fact that such rumours can prove to be more dangerous on a short period of time, and need to be identified as soon as possible in order to avoid their spread, especially if their intent is malicious.

2.3. Other

Aside from fake news and rumorus, that represent the main topics of interest for the present work, several other types of misinformation and false content on the web have been considered in the literature. In the following, we will provide a brief description for the most widely studied problems.

Clickbait refers to article titles or social media posts whose aim is to attract readers to follow a link to the actual article page [19]. In addition, clickbait headlines have been identified as one of the major contributors to the spread of fake news over the web [86].

Differently, *social spammers* on social media refer to users who coordinately launch various types of attacks, such as spreading viruses or ads, and phishing [85].

Related to social spamming, another widely studied aspect is that of *fake reviews*. Fake reviews are typically found on e-commerce websites and media review aggregators (e.g. Rotten Tomatoes[7]), and typically aim to improve or disrupt the popularity of a given product.

As previously stated, we will not focus on these specific types of misinformation. Our analysis will focus on the main issues of fake news and rumours detection.

3. Data Collection

The collection of relevant data for analysis on fake news and rumours on the web is one of the main issues in this area. Researchers have to face a series of problems. First, as we discussed in Section 2, they have to manage different types of false information in the context of web and social media platforms. For example, rumours on social networks, fake news articles on malicious websites, fake reviews, etc. Second, the amount of false information is a small fraction of online content produced every day, even if we restrict our focus on news articles and posts discussing breaking news. Third, social media companies have nowadays strict policies for what concerns the analysis of data produced by their users. This is especially true after the Facebook and Cambridge Analytica data scandal surfaced in the first months of 2018. Finally, given the different types of misinformation, several different tasks have been proposed in the literature, such as fake news detection, clickbait detection, rumour detection, and rumour veracity classification. For each task, different means of collecting and annotating data may be necessary. For these reasons, a few benchmark datasets and data repositories are today publicly available.

In this section, we will present an overview on data that can be employed to analyze the phenomena of fake news and rumours on web and social media, as well as an overview on the publicly available resources.

[7] https://www.rottentomatoes.com

3.1. Fake News

Fake news articles are mostly found on malicious websites, specifically created to spread misinformation. They are often later shared on social media platforms by authors, malicious users working with them or social media bots, as well as inattentive users who do not bother to check the source of the article before sharing it.

In fact, the most straightforward method aimed at determining a potential fake news article is to asses its source. Certain fake news websites are built to resemble proper news outlets, by mimicking both the visual aspect and the domain name. For example, *ABCnews.com.co* resembles *ABCnews.com*. This is done in order to deceive users and let them believe that they are browsing and sharing from a reliable source of information. Other sources of fake news can be websites promoting conspiracy theories, alternative facts, and alternative medicine.

For this reason, such websites can be used to harvest articles, which have a high probability of being false. However, we must note that inferring veracity of a piece of news solely based on its source could be misleading. Moreover, fake news may also be found on verified sources. This could happen for example by mistake, or for the rush of publishing breaking news without properly checking sources beforehand. Thus, it is clear that a proper annotation of data, to be conducted by professionals with knowledge on the matter and access to many different sources, is strongly advisable.

In [79], the authors have provided an interesting overview of key requirements for a reliable fake news detection corpus. Both truthful and deceptive news articles have to be collected. In addition, it is fundamental to verify the ground truth for each element in the corpus. For the authors, key factors are also homogeneity in length and writing matter (i.e. news genres and topics). Moreover, they suggest the importance of determining the manner of news delivery (e.g. humour, newsworthiness, satire, sensationalism, etc.) for contextualizing the piece of news.

In the process of gathering fake news, a key aspect to consider is how to perform fact checking in order to obtain a reliable data set of fake and truthful news. In [85] the authors have provided an overview of existing fact checking approaches in the context of *knowledge-based* fake news modeling and identification. *Expert-oriented fact checking* relies on human experts to assess the veracity of news. This process is exploited for example by websites such as *Snopes*[8] and *FactCheck.org*[9]. The main issue with expert-oriented fact checking is the cost both in time and resources. A possible alternative is to implement *crowdsourcing* in the fact-checking process, in order to exploit the "wisdom of the crowds" to annotate potentially suspicious news content, as in services such as *Fiskkit*[10]. Finally, [85] mentions *computational-oriented fact checking* models, that are often based on algorithms and external resources (e.g. knowledge graphs and the open web) to assess both check-worthy piece of news and their veracity. In particular, expert-oriented and crowdsourcing-oriented fact checking can be exploited to reliably annotate datasets of fake news.

In addition, the aforementioned fact checking services could be exploited to build reliable collections of fake news. In fact, websites such as Snopes provide fake news in the form of statements, that have either a true or false status. Moreover, often the source of the fake news is available, in the form of a social media post or fake news article.

Finally, humorous and satirical websites, which produce fake news, are also worth mentioning. They may be a useful source of content especially for tasks such as satire and irony detection. The most notorious example is *The Onion*[11], an American satirical website. Articles produced for these sites often rely on actual events and stories, but the content is clearly false or unbelievable. The authors' intention in this case is not to misinform but rather to amuse readers. However, since the format of links on social media is almost identical for any web source, careless users may believe and share such stories. Satirical websites

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

[9] https://www.factcheck.org

[10] http://www.fiskkit.com

[11] https://www.theonion.com

as a source of data have been exploited for example by [78] to perform *satirical news identification*.

3.2. Rumours

Rumours are mostly studied on social media. This is because social media platforms are often used to share information as quickly as possible between users. This may result in sharing unverified information that, in turn, may spread and generate a rumour.

The literature on how datasets should be collected for rumours detection and analysis mostly focuses on two main strategies, namely *top-down* and *bottom-up* collection strategies [111]. Top-down strategy requires some form of a-priori knowledge about target rumours. In particular, rumours are usually collected, after they spread on social media, by searching for a specific set of *keywords* and *tags* that describe the rumour. The proposed strategy is quite straightforward to implement and thus has been employed in several researches related to rumour detection and verification [15, 64, 75, 97, 109]. In this context rumour debunking websites are often used as source to identify the most interesting rumours and to extract reliable keywords for retrieving posts about those rumours on social media [22, 51, 57, 62, 63, 75].

Major downsides of top-down collection strategies are: i) rumours have to be known a-priori, making such datasets not suitable for the discovery of breaking news type of rumours, and ii) retrieving all social media posts related to a given rumour is problematic due to limitation on social media API services. However, a top-down approach could nonetheless be useful to collect data on long-standing rumours or controversial topics and themes that may spark different rumours, such as for example vaccines [25].

On the contrary, a bottom-up strategy is specifically aimed at collecting potential rumours in breaking news. This collection strategy has been proposed in [29, 36, 113, 116]. The main idea is to gather as many social media posts as possible during a certain time window. Then, the collected posts are evaluated by human annotators as a timeline of events in order to annotate them with various levels of analysis (e.g. rumour/non rumour, true/false, etc.). This strategy has a number of advantages over a top-down approach. However, it poses some problem that must be held into account. In particular, it is clearly more costly in terms of both human resources (i.e. the annotators) and computational resources. Moreover, for any given event, the potential loss of data in terms of related posts may be larger than in top-down strategy. Finally, it may also happen that during the collection stage a limited number of rumours of interest emerge [111].

In addition, one important aspect to take into account when collecting rumour data is the proposed aim of the analysis. Many subtasks in rumour analysis have been suggested, namely *rumour detection*, *rumour veracity classification*, *rumour stance classification*, and *rumour tracking* among others [111]. Each subtask may necessitate different types of data and different forms of annotation.

Rumour detection is the first step in a rumour classification system [111]. The proposed task is to classify a set of social media posts organized in a timeline as either *rumour* or *non rumour*. Researches in rumour detection have successfully applied both strategies to collect data. For instance, the authors in [63, 65, 75] have used a traditional top-down approach, employing keywords related to known rumours as well as extracted from rumour debunking services. On the contrary, [36, 113] have successfully employed a bottom-up strategy.

Rumour veracity classification can be considered as the other major subtask in rumour analysis, and has been extensively studied from various perspectives. Given an already detected rumour, i.e. a statement containing unverified information, and a set of posts (and respective metadata) associated with such rumour, the aim of veracity classification is to classify the rumour as *true*, *false*, or still *unverified*. In [113], the authors have proposed both a dataset and an annotation scheme for veracity classification, which has been exploited in the rumour veracity task at SemEval-2017 [29]. Many studies faced the problem employing a top-down strategy, by collecting data exploiting rumour debunking services (e.g. Snopes, PolitiFact and the Sina Weibo rumour debunking service) to identify true and false rumours, and then collect data directly from such services or identify related posts (e.g. via keywords) on social media [18, 22, 57, 62, 64, 76, 103, 105, 109]. A bottom-up oriented approach has been employed in [8, 36, 113]. In this case, breaking news related posts are collected and manually annotated by experts for their veracity. In addition, [95] has proposed the use of suspicious and trusted news account on social media to retrieve relevant posts (e.g. *retweets*). It is clear that access to trusted services such as rumour debunking websites is fundamental in order to identify true and false rumours.

Other tasks such as *Rumour Tracking* and *Rumour Stance Classification* have been addressed in the literature [111]. Collection of data for such tasks closely follows the strategies presented for detection and veracity classification. In the case of rumour tracking, the task is to annotate social media posts as either related or unrelated to a given rumour. Stance classification concerns instead the identification of opinions in social media posts with respect to a given rumour: the classification has been performed at the beginning as a 2-class problem (*support* and *deny*) [75], or more recently as a 4-class problem (*support*,*deny*, *question*, and *comment*) [74, 116]. Data for this task have been collected both using a top-down [74, 75], and a bottom-up strategy [113].

3.3. Publicly Available Datasets

Given the aforementioned difficulties in collecting relevant data, not many datasets have been collected and even fewer are publicly available.

3.3.1. Fake News

To the best of our knowledge, an agreed benchmark dataset for fake news detection has not been produced yet [85]. The main reasons are the difficulty of providing a clear definition of fake news (see Section 2), and the trouble of collecting relevant data for analysis. However, some publicly available resources are worth to be mentioned.

Many authors have focused on the creation of datasets containing statements from social media, for example made by politicians or public figures, labeled with information about their veracity. [94] has introduced the task of fact checking and produced a dataset of statements, containing both a veracity assessment and an analysis of the reason behind such assessment. Similarly, *LIAR* [101] contains short political statements, obtained through the website PolitiFact.com[12]. Each statement is annotated with the author, the context, a veracity label and a justification for such label. [89] has synthetically produced statements by altering Wikipedia sentences, and then providing evidence for or against such claim in Wikipedia articles.

For what concerns instead proper fake news, [35] has provided a dataset containing both rumoured claims and related news articles, annotated for their veracity.

The authors in [73] have focused on Facebook publishers. Starting from BuzzFeed[13], the collected data contain URLs from posts produced by nine verified Facebook publishers (3 mainstream publishers, and 6 hyperpartisan publishers). Each post is manually fact checked and annotated for veracity by BuzzFeed journalists. The interesting aspect of this dataset is that it considers fake news from a more social media-oriented perspective. [82] has subsequently added information regarding comments to this dataset. [88] has also collected Facebook posts from scientific and pseudo-scientific pages.

CREDBANK[68] is a large scale dataset, containing 60 million tweets. Tweets are grouped into events by means of topic modelling techniques. Each event is annotated for credibility via Mechanical Turk. Although a credibility judgment cannot be directly associated with a veracity score, it may be used as a good approximation for veracity classification.

From a non-research related perspective, it is interesting to cite BS Detec-

[12] http://www.politifact.com

[13] https://www.buzzfeed.com/

tor[14], developed by *Kaggle*[15]. It is a web crawler with knowledge about fake news websites, that has been used to build a dataset by monitoring such websites for a period of time. The main issue is that it cannot be considered a gold standard since it is not annotated by humans.

Finally, [84] has provided the most complete dataset of statements in terms of related information. Authors have built a system for fake news identification, and provided a dataset containing information about the content (textual and visual), as well as information about social context (i.e. users, network information, etc.) and characteristics of the spread evolution.

3.3.2. Rumours

As regards rumours, a more extensive effort has been undertaken in order to gather relevant data, especially in the context of the PHEME project [28].

The most relevant dataset, that can be considered as a benchmark for different possible evaluation purposes, has been collected by [116]. The dataset includes tweets related to 9 different rumours collected from 2014 to 2015. A bottom-up strategy was followed to gather data. Subsequently, tweets with a high retweet count were annotated on different levels (rumour/non rumour, true/false/unverified etc.) by a group of journalists. Moreover, tweets were grouped by events and *stories* within each event. Each story has information about the *resolving tweet*. In addition, responses to tweets were collected as well, and annotated for stance with respect to the source tweet. The dataset has beed used both in research [113] and for the RumourEval task in the SemEval 2017 evaluation campaign [29].

Finally, the previously mentioned CREEDBANK [68] may be a useful resource, especially concerning the veracity classification task of rumours on social media.

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

[15] https://www.kaggle.com/mrisdal/fake-news

4. Feature Extraction

False information detection in web-based communication has been approached from different perspectives, such as *Natural Language Processing* (NLP), *Data Mining* (DM), and *Social Network Analysis* (SNA). Detection methods will be discussed in section 5. In this section, we describe the information the different approaches consider as relevant for the analysis.

A very broad but interesting distinction in this regard has been proposed in [85]. Authors, focusing mostly on fake news, distinguish between *content-based* and *context-based* approaches to feature extraction. On the one hand, content-based approaches rely on *content features*, which refer to information that can be directly extracted from text, such as linguistic features. On the other hand, context-based approaches are more varied, and generally rely on surrounding information, such as user's characteristics, social network propagation features and reactions of other users to the news or post. Figure 4 summarizes a simple categorization of the different types of features that will be described in this section.

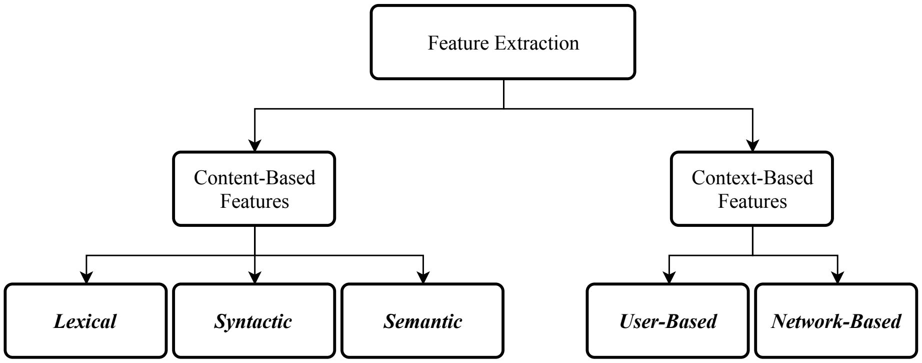

Figure 4: Different types of features used in the literature for fake news and rumour detection.

In addition, several current approaches tend to employ a mixture of content and contextual information to identify fake news and rumours. This is because it has proven challenging to completely automate fake news detection based only on a single type of features [81].

4.1. Content Features

Content features are generally directly extracted from texts. Most researches have exploited, at least in some capacity, such features. The obvious advantage is that, given the right tools for the analysis, textual features are readily available. NLP toolkits such as Stanford CoreNLP[16] and NLTK[17] can be, for example, implemented for extracting basic knowledge from text (Tokens, Sentences, Part-of-Speech (PoS), Lemmas, etc.) and obtain a structured representation, more suitable for further analysis.

A large body of work in linguistics and NLP exists concerning the identification of linguistic cues for detection. Earliest studies on computer mediated communication [13, 110] had already shown promising results. Moreover, works specifically focused on social media communication [12, 72, 108] have been able to isolate various linguistic cues which characterize deception, such as the use of self reference, negative words, swear words. The most recent studies have focused on both linguistic categories signaling credibility and specific phrases [69]. These linguistic cues, in addition to other features, have been successfully implemented in numerous approaches concerning fake news and rumour detection [15, 38, 39, 63, 76, 78, 95, 108].

We can distinguish among *syntactic*, *lexical* and *semantic* linguistic features.

Syntactic features, such as the number of content words (nouns, verbs, adjective), and the presence and frequency of specific POS patterns, have been employed, for instance, in [75, 76, 113]. In addition, sentence complexity has been identified as an indication for the reliability of the information [12, 97] as well as index of subjectivity [22, 40].

Lexical features concern actual word usage in texts. In particular, the most straightforward approach is to exploit the most salient content words or expressions (e.g. bi-grams and tri-grams) as features [16, 34, 73, 78]. Moreover, it has been proven that negations, doubt words, abbreviations and vulgar expressions

[16] https://stanfordnlp.github.io/CoreNLP/

[17] https://www.nltk.org

can be insightful in determining veracity of online information [97]. Finally, [76] has exploited features based on novelty of words found in social media posts.

Semantic features are often extracted by means of advanced NLP techniques. For example, *sentiment analysis* and *opinion mining* approaches adopt features based on the opinions and emotions expressed in the text [15, 39, 75, 105]. In addition, *topics* of social media posts, extracted by means of Latent Dirichlet Allocation (LDA) [7], have been proposed as features by [48]. *Distributional semantics* techniques such as *word embeddings* [67] have proven to be useful in numerous contexts throughout the NLP research field, and have been successfully implemented also for fake news and rumours detection, especially in machine learning and deep learning approaches [51, 63, 81, 113].

Content features have been extensively studied and implemented for fake news detection. Clearly, the content plays a key role in discovering potential deception. However, such features present some downsides. First, textual cues to deception may achieve limited generalization capability in a real word application system. Second, they may lose in descriptiveness as fake news are becoming more and more similar to proper news for what concerns the writing style. Finally, especially concerning rumours on social media, given the relative brevity of texts, other characteristics may prove to be more effective.

4.2. Context Features

Context features are extracted by considering relevant information surrounding the actual social media post or fake news. In particular, the most used context features concern the analysis of users, sources of the rumour or news, propagation structures of the information on social media, and reaction of other users with respect to the news.

User-based features model characteristics of social media users (e.g., number of posts, age of the account, number of friends/followers, etc.) who produce or share the news. Earliest studies on user credibility [15] have adopted some features directly extracted from information on the users in Twitter, obtaining promising results. Moreover, user-based features have also been employed in categorizing users as genuine or fake (e.g., social spammers, bots, etc.) [47]. In most social media platforms, it is possible to obtain a *verified* status for the account. It is often used by celebrities as well as public figures in order to discourage the emergence of fake profiles. Clearly, it can also be considered as a good measure of the reliability of the shared information. In fact, most approaches employing user-based features include it in their analysis. Other relevant information concerns the amount of activity and social circles within the social media. This can be calculated for example as the number of posts as well as the numbers of *followers* and *followings* (or a follower to following ratio) [57, 113]. Moreover, many studies consider the age of the account on the platform, the presence of a clear description for the profile and of URLs linking to external resources [15, 17, 62, 64, 103, 105, 113].

Obviously, when considering user-based features, an important aspect to be taken into account is availability. In fact, most researches show similar trends for what concerns such features. This is mainly due to the fact that information on the users and user interactions on the platform is not generally accessible, due to privacy constraints. Further, only a handful of information is represented in the same way across different social media platforms.

Network-based features are used to model information concerning the properties of the network where the news is shared. For example, *propagation structures*, *diffusion patterns*, and properties of the *sub-graph* in which the news is spread, such as *density* and *clustering coefficient*, are taken into account.

According to [65], most existing studies implementing network-oriented features are limited to the use of statistics on the diffusion patterns, such as number of *retweets* and propagation times [15, 39, 64, 105]. Other studies have focused instead on modeling the temporal characteristics of propagation [56, 57]. For example, [57] has built several different networks, based on friendship status among users as well as propagation patterns of news, and extracted features based on clustering coefficient and degree of such networks. Finally, other works have incorporated both content and user based features as well as information from the propagation tree [97, 100, 103]. [65] is one of the first approaches to directly evaluate the similarity of propagation trees in order to infer the veracity of the source post.

In addition, some approaches have focused on exploiting the *wisdom of the crowds* to identify rumours and fake news. Such approaches in fact mostly implement the *stance* of other users in a platform towards a given rumour as features to determine its veracity. [116] has shown how an aggregate stance (e.g., *mostly positive* or *mostly negative*) may correspond to the veracity of a rumour. Albeit stance detection approaches have been proposed in the literature [29, 39, 75, 113], not many rumour or fake news detection systems, which employ such stance as feature, exist. [50], following [92], has exploited topic models [7] to identify conflicting viewpoints in microblogs, and has built a credibility network to determine the veracity of social media posts. [88] has used instead Facebook likes (i.e. users who like news posts) as features for classifying posts as either *hoax* or *non hoax*.

5. Approaches to Detect False Information

Most of the approaches proposed in the literature to detect false information face the task as a classification problem: they aim to associate labels such as *rumour* or *non rumour*, *true* or *false* with a particular piece of text. In most of the cases, researchers have employed machine learning and deep learning approaches, achieving promising results. Alternatively, some researchers have applied other approaches based, for instance, on data mining techniques, such as time series analysis, and have exploited external resources (e.g. knowledge bases), to predict either the class of documents or events, or to asses their credibility.

Most approaches aimed at detecting fake news have focused on using content features for classification. In fact, according to [85], very few approaches for fake news detection have relied on purely social-context models. On the contrary, rumour detection and verification approaches often use a mixture of content and context features for their models. Clearly, this is due to the fact that the social aspect of rumours may play a key role in improving detection performance.

As we pointed out in Section 3, there does not exist a widely studied benchmark dataset, especially concerning the fake news problem. Thus, considering also the number of different sub-tasks proposed in the literature, it is very hard to perform a reliable and fair evaluation of the false information detection approaches and compare their performance. Moreover, often researchers are forced to test different classification algorithms in order to find the most suitable for the available dataset and task. Figure 5 summarizes the different approaches we will consider in our analysis. These approaches can coarsely be grouped into classification approaches and other approaches. The classification approaches can be in their turns divided into approaches based on machine learning and deep learning. The following subsections will discuss in detail approaches based on machine learning and deep learning, and other approaches.

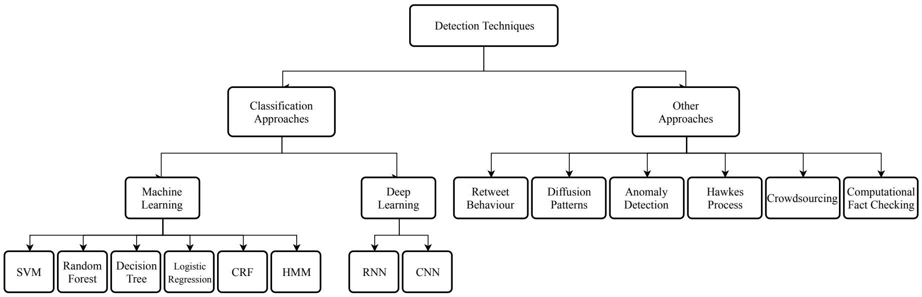

Figure 5: Different approaches to fake news and rumour detection proposed in the literature.

5.1. Detection Approaches based on Machine Learning

Machine learning algorithms have proven to be extremely useful for solving numerous tasks in the information engineering field. Since the earliest approaches focusing on credibility on social media [15] and deception detection in computer mediated communication [108] had provided promising results, machine learning techniques have been implemented in a number of researches concerning the problem of false information detection on the web. In particular, most machine learning approaches implemented for fake news and rumour detection have employed a supervised learning strategy.

Support Vector Machines (SVMs) are one of the most widely used methods for classification in a number of research areas. SVMs are discriminative classifiers formally defined by a separating hyperplane.

According to the experiments in [108], SVMs have outperformed a number of supervised machine learning approaches for deception detection in text, obtaining an F-measure F_1 of 0.84. However, as pointed out by the authors themselves, there exists a significant variation in performance depending on the dataset selected for training [108].

Content-based features (e.g. linguistic and visual features) were exploited in most SVM-based approaches to fake news and deception detection [1, 12, 72, 78, 108]. In particular, [1] has obtained highly competitive scores for the task of deception detection on a number of datasets by exploiting only lexical, syntactic, and content-specific features. [78] has trained an SVM for satirical fake news detection with a number of content-based features, obtaining an F_1 of 0.87.

As regards rumour related tasks (i.e. detection, verification etc.), SVM-based approaches have made a more prominent use of context-based features as well as content-based ones. Most approaches have exploited a mixture of content- and context-based features to train the classifiers, which are used to tag either a single social media post or an event described in a series of posts as rumors or not [76, 105]. Both [105] and [76] have proposed an extension of the feature set presented in [15]. Further, [76] has also presented a set of features aimed at detecting novelty in tweets, that has allowed achieving an accuracy of 0.75.

Another interesting SVM-based strategy has been employed in [103], where the authors have proposed the use of a graph-kernel-based SVM classifier to identify rumours using propagation structures and content features. An accuracy of 0.91 was reported on a small Sina Weibo dataset.

In [45] authors have proposed a set of features to distinguish among fake news, real news and satire. The experimental results reported in the paper suggest that titles are crucial in distinguishing between real and fake news, and that content of fake news is more akin to satire.

SVMs were also exploited for the task of clickbait detection in [16]. Authors used a set of content-based features to obtain an F_1 of 0.93.

Another widely studied family of algorithms, proposed particularly for rumour analysis tasks, is *decision tree* [11]. Decision tree performs a recursive split on feature values in order to determine the class. Decision trees are generated from data with algorithms such as J48 (C4.5) [77]. Despite their relative simplicity with respect to other machine learning schemes, they exhibit competitive performance on the task at hand.

The effectiveness of *J48 decision tree* with respect to other algorithms, including SVM, has been shown in [3, 15, 36, 109]. More specifically, in [15] the authors have used a mixture of content-based and context-based features to perform credibility evaluation of tweets, achieving an accuracy of 0.86. In [8] the authors have proposed a series of user trust metrics in order to evaluate the trustworthiness of users in social media via decision tree. They have reported an accuracy of 0.75. Enquiring tweets were exploited in [109] in order to determine clusters of potential rumours. For each cluster, a series of features describing it are extracted and fed to a J48 decision tree. Authors report an accuracy of 0.52 for their best run. [36] has implemented a decision tree for veracity classification of rumours. The approach is able to achieve 0.96 accuracy on the proposed dataset, with a small number of features.

Random forests [10] have been employed in a number of works on rumour analysis. A random forest is an ensemble of several decision trees. The mode of all the predictions from each tree is the final output. Comparative studies on various machine learning algorithms for rumours and fake news tasks have outlined random forests as a strong performer [3, 12, 36, 57, 107]. More specifically, [57] has implemented a random forest with a set of temporal, structural and linguistic features for rumour classification in a tweet graph, obtaining an accuracy of 0.90. Random forest is also exploited for stance detection in [3, 107]. [107] has used only content-based features of tweets and has obtained an average accuracy between 0.83 and 0.88 on a manually collected dataset. In [3] the authors have proposed the use of a set of features derived from previous works [39, 62, 75, 116] as well as a set of stance-specific features. They have reported 0.79 accuracy on the RumorEval dataset [29]. Random forest was also exploited for performing user credibility evaluation in [48]. Authors have proposed the extension of traditional feature sets for user credibility assessment, like in [15], with topic-models based features extracted by means of Latent Dirchlet Allocation (LDA)[7].

Learning algorithms based on *logistic regression* models have been also employed in a number of studies, in particular for rumour stance classification tasks. As for decision trees and random forests, studies comparing different approaches in the context of rumours and fake news have reported competitive performances for logistic regression [3, 32, 36, 88, 107, 110].

Logistic regression was used for stance classification of news articles based on headlines and claims in [35]. The proposed method has obtained an accuracy of 0.73 on the Emergent dataset [35, 86]. [22] has performed a study on linguistic predictors of rumor veracity by exploiting a logistic regression to identify the most significant ones with respect to true and false rumours. [40] has exploited logistic regression in analyzing the credibility of Bulgarian true and false news, achieving an accuracy of 0.75 on the hardest presented dataset.

Conditional Random Field (CRF) classifiers [58] have received less attention for solving tasks in this domain, despite their relative popularity in NLP tasks such as for example PoS-Tagging. The main advantage of a CRF classifier is the capability of modelling sequences. [113, 114] has argued that a CRF is able to leverage previous context in order to improve the detection of rumors for breaking news events. Authors have reported an F_1 of 0.6 on the PHMEME dataset [113, 116]. Results have shown that CRF outperforms non-sequential approaches as well.

Hidden Markov Models (HMMs) are capable of dealing with sequence-based data too. HMMs have been exploited for rumour veracity classification in [96, 97]. Authors have proposed to model rumour features, both content- and context-based, in time-series form. Time series of observed data have been then used to train two HMMs, one for true and one for false rumours. The likelihood of a new rumour is then estimated with both models, and the one with the highest likelihood is chosen. The proposed approach has obtained an overall accuracy of 0.75 on the proposed dataset.

Finally, an approach based on ensemble of classifiers has been implemented in [98]. Ensembles rely on the predictions of a number of independent classifiers, and provide the final classification via a majority vote scheme. [101] has proposed the use of a series of content- and context-based features fed to several learning algorithms. The ensemble is able to achieve 0.77 accuracy on the RumourEval test dataset [29].

5.1.1. Detection approaches based on Deep Learning

Deep Learning is one of the most widely explored research topics in machine learning. Deep Learning classifiers have seen an unprecedented rise in popularity in recent years, due to extremely promising results in a number of research fields, including text mining and NLP. Deep learning frameworks have a main advantage over traditional machine learning approaches. Indeed, traditional machine learning representations are based on manually crafted features. The feature extraction task is time-consuming and may result in biased features [63]. This is a critical issue for tasks such as fake news and rumour detection, where the identification of relevant features for the analysis may pose an even greater challenge. On the other hand, deep learning frameworks can learn hidden representations from simpler inputs both in context and content variations [63]. The problem is therefore shifted from modeling relevant input features to modeling the network itself in a way that enables the task to be solved efficiently.

The two most widely implemented paradigms in modern artificial neural networks are *Recurrent Neural Networks* (RNN) and *Convolutional Neural Networks* (CNN).

RNNs are a class of neural networks in which nodes are connected sequentially to one another to form a directed graph. This structure makes RNNs, and in particular Long Short-Term Memory (LSTM) networks [44], particularly effective for modelling sequential data, such as for example human language, and capturing relevant features from different sources of information [81]. In the context of fake news and rumours, the earliest adoption of RNNs for rumour detection is reported in [63]. Authors have proposed different RNN architectures, namely *tanh-RNN*, LSTM and Gated Recurrent Unit (GRU) [21]. Among the proposed architectures, GRU has obtained the best results in both the datasets considered, with 0.88 and 0.91 accuracy, respectively. In addition, all deep learning architectures were shown to outperform traditional machine learning algorithms. An LSTM network as part of a framework for fake news detection has been employed in [81]. The LSTM is fed with temporal data concerning the engagement of a news, user properties and text properties. The approach has been tested on the datasets proposed in [63], obtaining an accuracy of 0.89 and 0.95, respectively. In [112], different approaches based on *sequential models* for stance classification have been evaluated, namely Linear and Tree CRF, Hawkes Process, and LSTM. In particular, authors have exploited features extracted from the sequential interactions of users on Twitter to classify the stance of each tweet. LSTM has outperformed all the other approaches. Finally, the authors in [54] have proposed a *multi-task learning* approach to solve the problem of rumour classification. More specifically, they have designed a multi-task learning framework with an LSTM layer shared among all tasks, as well as a number of task specific layers. Performance have been evaluated on the RumorEval [29] and PHEME [113, 116] datasets. Authors have reported a per-event accuracy ranging from 0.36 to 0.64.

CNNs are a class of neural networks with an input layer, an output layer, and a series of hidden layers, where a number of transformations are applied to the data by means of pooling and convolution operations. CNNs have been widely studied for image recognition and processing [60] and are considered the state-of-the-art for many *computer vision* tasks. However, CNNs are gaining popularity in the NLP community as well [49]. Considering rumours and fake news, a number of recent works include CNNs. [20] has proposed a CNN with single and multi-word embeddings for solving both stance and veracity classification of tweets. Authors have reported an accuracy of 0.70 for the task of stance classification, and of 0.53 for the task of veracity classification in the RumorEval evaluation campaign [29]. CNNs and paragraph embeddings [59] have been exploited in [106]. Authors have proposed the use of paragraph embeddings to learn representations of small groups of posts in a specific event and have used them as input for their CNN model. The approach has been evaluated on the datasets proposed in [63], obtaining an accuracy of 0.93 for Sina Weibo and 0.77 for Twitter. [95] has evaluated both an RNN and a CNN approach for the identification of suspicious (e.g., propaganda, hoaxes) and trusted news posts. The proposed architecture implements both word sequences and linguistic and network cues of deception. Both approaches clearly outperform baselines and obtain an average precision close to 1.00.

In addition, a number of recent works have exploited a mixture of RNNs and CNNs in their models [2, 87, 101]. [101] has proposed a hybrid approach to fake news detection on the LIAR dataset [101]. The proposed architecture encodes text information via a CNN, and metadata about the author of the text by means of an LSTM layer. The hybrid model has proved to outperform all baselines, including a bi-LSTM architecture, with an accuracy of 0.27 on a portion of the dataset used as test. [2] has performed some experiments using LSTM and hybrid LSTM-CNN architectures. The best results were obtained by the simplest LSTM model, with an accuracy of 0.82 on the PHEME dataset [113, 116]. However, the authors themselves have stated that the hybrid model is expected to obtain better results given a larger training set. [87] has proposed to detect false rumours on social media based on repost sequence patterns. More specifically, authors have exploited a CNN to extract feature vectors from posts and reposts, and have used them as inputs to an RNN. In addition, authors have tried to perform detection as early as possible in the sequence of reposts based on a threshold strategy. The approach has been tested on the datasets in [63], obtaining accuracies above 0.90.

5.2. Other Approaches

Aside from machine and deep learning-based classification approaches, other studies have experimented with a number of different approaches to fake news and rumours detection. The importance of such approaches is twofold. First, they can represent valuable alternatives to machine learning and, more in general, classification approaches, by possibly exploiting different characteristics of false information. Second, they spark research efforts in different directions and involve a broader and more interdisciplinary research community around the topic of interest in this paper.

The approaches discussed in the following tend to exploit models, such as clustering and vector space models, different from the ones presented above in order to identify fake news, rumours and their properties.

A framework for evaluating tweets and their authors as possible sources of misinformation has been proposed in [55]. In particular, the credibility of a tweet was first evaluated by measuring disparity in retweet behaviour via the Gini coefficient. Credible tweets and their retweets are then used to build a retweet graph. The PageRank algorithm [71] was then used to compute a score of acceptability for the given tweet. Interestingly, such information is then presented to users with the objective of allowing them to perform a more informed evaluation of the tweet.

A study on the analysis of patterns of diffusion to detect rumours has been proposed in [99, 100]. The authors have identified a series of short diffusion patterns, based on stance, that appear to be strongly related to rumours [100]. In [99], such patterns have been implemented in a sliding-window based framework to detect rumour events as early as possible by exploiting frequent pattern mining techniques. Authors have reported an accuracy of 0.70 with their approach.

A more linguistic-oriented perspective has been explored in [80] to detect deception in news by exploiting *Rhetorical Structure Theory* (RST) to describe documents. The aim is to represent truthful and deceptive news in terms of coherence by means of RST. After the RST analysis, Vector Space Models (VSM) are used to describe each document. Starting from a manually annotated dataset, authors have built a VSM containing two centroids for truthful and deceptive news. New documents are compared by exploiting these centroids. Authors have reported an agreement of 0.67 between human assessments and the proposed method.

False rumour detection was managed as an anomaly detection problem in [18]. In particular, the use of an extension of Principal Component Analysis (PCA), namely *Factor Analysis of Mixed Data* (FAMD) [61], as means of anomaly identification was investigated. First, a set of features for each social media post were identified. Then, FAMD was used to reduce the dimension of the feature space. Finally, rumours were ranked by their deviation degree by exploiting a combination of cosine similarity and Euclidean distance. The proposed approach has obtained F_1 of 0.79 and 0.81 for detecting rumours and non-rumours on a Sina Weibo dataset, respectively.

In [33], the authors have explored the problem of mitigating the effect and spread of fake news on social networks. They have formulated the problem of fake news mitigation as an optimal intervention strategy enacted by users, denoted as *mitigators*, in the network where the news is spreading. To this aim, *Multivariate Hawkes Process* [42] and *Least Squares Temporal Difference* [9] have been employed to define optimal procedures for mitigators.

In [88], algorithms derived from *crowdsourcing* have been used for classification of Facebook posts as hoaxes or non hoaxes. The main idea of the proposed method is that hoaxes can be identified by the users who interact with them. An adaptation of the algorithm proposed in [26] has been used in order to learn Facebook users tendency to like either hoaxes or non hoaxes. The model has been then tested on the proposed dataset, obtaining 0.99 accuracy. In addition, [27] has proposed to incorporate also content-based features to the method presented in [88], outperforming previous results, and obtaining an accuracy of 0.81 on a real world dataset of Facebook posts with fake news articles.

A number of recent approaches have exploited the idea of tensor decomposition in order to perform unsupervised or semi-supervised fake news detection [37, 46]. In particular, [46] has proposed to decompose news articles into tensors able to model spatial relations between terms in a document, via CP/PARAFAC decomposition [41], and then perform co-clustering on such tensors in order to identify similar documents and outliers representing fake news. Similarly, the authors in [37] have performed tensor decomposition of news articles and used such tensors to build a graph of k-nearest neighbors articles. Then, given a small set of labeled articles, they have treated the problem as a semi-supervised learning problem over graph, by exploiting a belief propagation algorithm. Authors have tested their approach on a number of datasets, obtaining an accuracy of 0.69 using only 5% of labeled data, and 0.73 using 30% of labeled data.

Finally, approaches based on *computational-oriented fact checking* [85] are worth to be mentioned. Although the task of fact checking can be considered as slightly different from fake news and rumour detection, it is nonetheless definitely akin to them. The main effort is addressed to perform automatically fact checking. A number of strategies have been proposed to this aim. [66] has automated the process of web-based fact checking, by comparing facts extracted from a given document against facts extracted from URLs related to such documents. [104] has introduced the task of automated fact-checking and presented a series of algorithms for solving the task by automatically designing queries aimed at checking whether the statement is true or false. The most widely used technique is however the exploitation of knowledge graphs. Such graphs have been employed in a number of studies [23, 83]. More specifically, the use of Wikipedia *info-boxes* to generate a knowledge graph has been proposed in [23]. Here, the authors have defined a measure of semantic proximity by using a transitive closure algorithm in order to check claims against the knowledge graph. In [83] the problem has been tackled as a link prediction task in a knowledge graph. Each statement corresponds to a path in the graph: the existence of meta-paths is exploited with the aim of reducing the search space with respect to the statement's path.

The overview of the different fake news and rumours detection works presented in this section has highlighted that several different approaches have been proposed in the literature, exploiting various techniques. The performances, expressed in terms of accuracy or F-measure, of these approaches are quite different. However, since there do not exist benchmark datasets, it is practically impossible to compare the different approaches and evaluate whether one approach is more performing than another. Nevertheless, we have decided to report the performance presented in the papers to allow the reader to have a glimpse of what they can expect at the state of the art in terms of performance for this type of application domain.

6. Future Directions

The issues discussed in this paper are definitely relevant, from both real-world and research standpoints. The topics of fake news and rumours on the web are relatively new. However, our review of the literature has shown that they are a thriving field of research, and results obtained from both the research and application perspectives are indeed very promising. We expect in fact that previously discussed approaches and techniques will be further improved and will have the potential to be implemented into real-world applications to fight the spread of rumours and fake news online. However, it is clear that improvement is still needed concerning many aspects of the discussed problems.

We will now briefly discuss some of the major challenges in fake news and rumours detection and will provide insight on possible future directions for research.

First and foremost, the lack of widely accepted benchmark datasets, especially concerning fake news and associated social media posts, has to be addressed. This is fundamental to evaluate the effectiveness of each approach and compare the approaches among them. In fact, available resources may not be sufficient for: i) gaining novel insight on relevant properties of fake news and rumours and ii) building models able to properly operate in a real world scenario. The most promising efforts on data gathering have been made by [116], concerning social media posts and discussions during breaking news events, and by [84], that has collected true and false statements from news and social media with a number of other context-based features. Research on collecting data for this kind of analysis should focus on building large scale datasets and most importantly on the identification of a clear and accepted benchmark for evaluation. Large scale datasets could enable analysis on a level more similar to real scenarios. Moreover, given the problem of clearly identifying relevant features for the analysis, as much information as possible should be collected in order to gain deeper insight on the problem.

Our evaluation of the literature has shown a clear trend in favor of supervised classification approaches to fake news and rumours detection. The trend will probably continue in future years, as supervised classification models based on deep learning keep growing in popularity among researchers for a wide variety of tasks. In fact, deep learning techniques have obtained state-of-the-art results, and most recent approaches are focused on exploiting such frameworks to some extent. This is probably due to the fact that, unlike traditional machine learning approaches, no effort required by feature engineering is needed to obtain a reliable prediction. Feature engineering is in fact an area where we can identify vast room for improvement, for two main reasons. First, fake news and rumours still have to be fully understood from a linguistic perspective [85], and research around telltale linguistic-based features has the potential of making several steps forward. Second, given the *streaming* nature of social media, information could be subject to *concept drift* [102]. This means that the relevance of data, as well as proposed features, could vary over time. Both content-based and context-based features, extracted for a given event or time frame, may not scale and generalize well for real-world applications or data from a different context. For these reasons, future research in this direction should focus on better understanding the importance of certain features for classification, as well as their ability to generalize on the problem and possibly manage concept drift in a real-world scenario. Moreover, the use of visual features has not received much attention in the literature. However, as photo and video manipulation tools become available to wider audiences, visual features to distinguish between true and fake content are increasingly important and could be studied more in depth.

Given our evaluation of the available literature, we can argue that the most effective approaches implement both content- and context-based features for classification. This has become apparent especially in rumour related research, where we can clearly identify a trend to use as much information as possible for detection. Hybrid approaches, which can simultaneously model different aspects of fake news, such as the actual text, diffusion patterns and stance towards the news, despite being more complex in terms of models, data availability and number of features, may be suited to better solve the problem and could be definitely pursued.

From the perspective of *models*, it may be beneficial to direct research efforts towards the implementation of *semi-supervised* or *unsupervised* models as a feasible solution. The main advantage of exploiting such techniques is given by the fact that they can learn from unlabeled data, thus mitigating the problem of finding and labelling relevant data. Moreover, such models could also enable a better understanding of the problem and of its key characteristics.

For what concerns alternative solutions to standard classification approaches, literature is still scarce. Nonetheless we believe that such approaches may provide crucial insights into the problem and should be further studied. Modeling the problem from a different perspective may allow to overcome the limitations of the proposed classification approaches, namely the need for labeled training data and potential lack of generalization capabilities in a real world scenario. Previously mentioned unsupervised or semi-supervised approaches, as well as knowledge-driven methods for automated fact checking [23, 83, 104] have shown promising results. For this reason, we believe that further studies focused on automatically modeling and exploring knowledge bases in order to find contrasting viewpoints, as well as automatically discovering patterns and regularities from unlabeled data, could improve detection of fake news and rumours.

Moreover, going forward in this field of research, several important aspects concerning the application of fake news and rumour detection methods should be taken into account and discussed among researchers. Despite the undoubted importance of tools able to help users of social media, and more generally of the web, there are a few concerning aspects that need to be addressed. First of all, such systems, in the wrong hands, may generate a form of automatic censorship, by means of producing false positives in the search for fake news. Independent sources of news may be targeted because of contrasting viewpoints with mainstream media. Second, news deemed as fake may turn out as real also months, if not years, later. A viable strategy to overcome these potential limitations could be to build real-world applications with such issues in mind. For example, the final decision of whether to trust the piece of news/rumour and the detection system could be left to the end user. Moreover, as earlier research has suggested [15], a continuous score of credibility or trustworthiness could be presented to the user instead of a dichotomous classification between true and fake.

Finally, such real-world implementations and tools have been mostly overlooked by the research community. Some effort in this direction has been described in [33]. It is clear that incorporating results in real-world systems could be advantageous for a number of reasons. First, enabling users with tools that can automatically validate the information as reliable may result in a drastic reduction in the sharing of fake claims. Second, feedback of users on such matter may prove extremely helpful for the improvement of models and strategies to spot false claims.

7. Conclusions

In this paper, we have explored the topic of potentially false information in computer mediated communication, and techniques for automatic detection of such information.

Fake news and rumours have become an integral aspect of our digital lives. They have already proven to be potentially very dangerous in the digital ecosystem as well as outside of it.

In this paper, we have provided an overview on current state of the art techniques and approaches to the problem of detecting such false information on the web. Our review has mainly focused on five key aspects. First, we have provided clear definitions and distinctions for the different sub-problems such as fake news and rumours. Research on such topics has in fact produced several different, albeit similar, definitions. We have tried to make the needed clarity in the debate on fake news and rumours. Second, we have highlighted valuable sources of relevant data and techniques to collect them. Since the field is quite novel, no widely accepted method for retrieving and labelling data has been proposed. Thus, several researchers face the same problem with slightly different goals. Although a considerable effort has been undertaken in the direction of publicly available resources, a widely accepted benchmark dataset has not emerged yet. Third, we have focused on the different aspects of fake news and rumours that can be implemented for detection, such as content-based and context-based features. Fourth, we have provided an overview of techniques to perform fake news and rumour detection and classification. Our review highlights how the main research trend is that of implementing supervised machine learning techniques. More specifically, while earlier research exploited traditional machine learning algorithms such as SVM and decision trees, most recent approaches rely on deep learning classifiers. Deep learning techniques have obtained state-of-the-art accuracy in most, if not all, instances. In addition, we have provided a description of the most promising alternative techniques, that either use unsupervised or semi-supervised learning algorithms or implement completely different strategies to perform the analysis.

Finally, we have provided some insight on possible future trends for this area of research. Such trends include improvement of existing methods by identifying novel relevant features for analysis as well as the identification of new alternative techniques that may better perform in a real world scenario.

Topics presented in this review are expected to become prominent in the discussion around social media, both from a social and a research standpoints. For this reason, we believe that further research in this direction is needed, as such contributions will definitely play a crucial role in shaping the future of online communication.

Acknowledgements

This work was partially supported by the University of Pisa [grant number PRA_2017_37] in the context of the project “IoT e Big Data: metodologie e tecnologie per la raccolta e l’elaborazione di grosse moli di dati”, by Tuscany Region in the context of the projects “Talent” and “Sibilla” in the framework of regional program FESR 2014-2020”, and by the MIT-UNIPI Project “Event Extraction for Fake News Detection”.

References

[1] Afroz, S., Brennan, M., Greenstadt, R., 2012. Detecting hoaxes, frauds, and deception in writing style online. In: Proceedings of the 2012 IEEE Symposium on Security and Privacy. SP ’12. IEEE Computer Society, Washington, DC, USA, pp. 461–475.

[2] Ajao, O., Bhowmik, D., Zargari, S., 2018. Fake news identification on twitter with hybrid cnn and rnn models. In: Proceedings of the 9th International Conference on Social Media and Society. SMSociety ’18. ACM, New York, NY, USA, pp. 226–230.

[3] Aker, A., Derczynski, L., Bontcheva, K., 2017. Simple open stance classification for rumour analysis. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017. pp. 31–39.

[4] Allcott, H., Gentzkow, M., 2017. Social media and fake news in the 2016 election. Tech. rep., National Bureau of Economic Research.

[5] Allport, G. W., Postman, L., 1946. An analysis of rumor. Public Opinion Quarterly 10 (4), 501–517.

[6] Allport, G. W., Postman, L., 1947. The psychology of rumor. The ANNALS of the American Academy of Political and Social Science 257 (1), 240–241.

[7] Blei, D. M., Ng, A. Y., Jordan, M. I., Mar. 2003. Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022.

[8] Bodnar, T., Tucker, C., Hopkinson, K., Bilén, S. G., 2014. Increasing the veracity of event detection on social media networks through user trust modeling. In: Proceedings of the 2014 IEEE International Conference on Big Data (Big Data). IEEE, pp. 636–643.

[9] Bradtke, S. J., Barto, A. G., 1996. Linear least-squares algorithms for temporal difference learning. Machine Learning 22 (1), 33–57.

[10] Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5–32.

[11] Breiman, L., 2017. Classification and regression trees. Routledge.

[12] Briscoe, E. J., Appling, D. S., Hayes, H., 2014. Cues to deception in social media communications. In: Proceedings of the 2014 47th Hawaii International Conference on System Sciences (HICSS). IEEE, pp. 1435–1443.

[13] Burgoon, J. K., Blair, J. P., Qin, T., Nunamaker, J. F., 2003. Detecting deception through linguistic analysis. In: Chen, H., Miranda, R., Zeng, D. D., Demchak, C., Schroeder, J., Madhusudan, T. (Eds.), Intelligence and Security Informatics. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 91–101.

[14] Cai, G., Wu, H., Lv, R., 2014. Rumors detection in chinese via crowd responses. In: Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, Beijing, China, pp. 912–917.

[15] Castillo, C., Mendoza, M., Poblete, B., 2011. Information credibility on twitter. In: Proceedings of the 20th international conference on World Wide Web. ACM, Hyderabad, India, pp. 675–684.

[16] Chakraborty, A., Paranjape, B., Kakarla, S., Ganguly, N., 2016. Stop clickbait: Detecting and preventing clickbaits in online news media. In: Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, pp. 9–16.

[17] Chang, C., Zhang, Y., Szabo, C., Sheng, Q. Z., 2016. Extreme user and political rumor detection on twitter. In: Proceedings of the 12th International Conference on Advanced Data Mining and Applications(ADMA). Springer, pp. 751–763.

[18] Chen, W., Yeo, C. K., Lau, C. T., Lee, B. S., 2016. Behavior deviation: An anomaly detection view of rumor preemption. In: Proceedings of the 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). IEEE, pp. 1–7.

[19] Chen, Y., Conroy, N. J., Rubin, V. L., 2015. Misleading online content: Recognizing clickbait as false news. In: Proceedings of the 2015 ACM Workshop on Multimodal Deception Detection. ACM, Seattle, WA, USA, pp. 15–19.

[20] Chen, Y.-C., Liu, Z.-Y., Kao, H.-Y., 2017. Ikm at semeval-2017 task 8: Convolutional neural networks for stance detection and rumor verification. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 465–469.

[21] Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, pp. 1724–1734.

[22] Chua, A. Y. K., Banerjee, S., 2016. Linguistic predictors of rumor veracity on the internet. In: Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS). pp. 387–391.

[23] Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., Flammini, A., 2015. Computational fact checking from knowledge networks. PloS one 10 (6).

[24] Conroy, N. J., Rubin, V. L., Chen, Y., 2015. Automatic deception detection: Methods for finding fake news. Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community 52 (1), 1–4.

[25] D'Andrea, E., Ducange, P., Bechini, A., Renda, A., Marcelloni, F., 2019. Monitoring the public opinion about the vaccination topic from tweets analysis. Expert Systems with Applications 116, 209–226.

[26] De Alfaro, L., Polychronopoulos, V., Shavlovsky, M., 2015. Reliable aggregation of boolean crowdsourced tasks. In: Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing. pp. 42–51.

[27] Della Vedova, M., Tacchini, E., Moret, S., Ballarin, G., DiPierro, M., de Alfaro, L., 2018. Automatic online fake news detection combining content and social signals. In: Proceedings of the 22st Conference of Open Innovations Association FRUCT. FRUCT'22. FRUCT Oy, Helsinki, Finland, Finland, pp. 38:272–38:279.

[28] Derczynski, L., Bontcheva, K., 2014. Pheme: Veracity in digital social networks. In: Proceedings of the 10th Joint ACL ISO Workshop on Interoperable Semantic Annotation (ISA). Reykjavik, Iceland, pp. 19–22.

[29] Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., Wong Sak Hoi, G., Zubiaga, A., 2017. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. In: Proceedings of the 11th

International Workshop on Semantic Evaluation (SemEval-2017). Association for Computational Linguistics, pp. 69–76.

[30] Di Fonzo, N., Bordia, P., 2007. Rumor, gossip and urban legends. Diogenes 54 (1), 19–35.

[31] Diakopoulos, N., De Choudhury, M., Naaman, M., 2012. Finding and assessing social media information sources in the context of journalism. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 2451–2460.

[32] Enayet, O., El-Beltagy, S. R., 2017. Niletmrg at semeval-2017 task 8: Determining rumour and veracity support for rumours on twitter. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 470–474.

[33] Farajtabar, M., Yang, J., Ye, X., Xu, H., Trivedi, R., Khalil, E., Li, S., Song, L., Zha, H., 06–11 Aug 2017. Fake news mitigation via point process based intervention. In: Proceedings of the 34th International Conference on Machine Learning. Vol. 70 of Proceedings of Machine Learning Research. PMLR, International Convention Centre, Sydney, Australia, pp. 1097–1106.

[34] Feng, V. W., Hirst, G., 2013. Detecting deceptive opinions with profile compatibility. In: Proceedings of the 6th International Joint Conference on Natural Language Processing. pp. 338–346.

[35] Ferreira, W., Vlachos, A., 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. pp. 1163–1168.

[36] Giasemidis, G., Singleton, C., Agrafiotis, I., Nurse, J. R., Pilgrim, A., Willis, C., Greetham, D. V., 2016. Determining the veracity of rumours

on twitter. In: International Conference on Social Informatics. Springer, pp. 185–205.

[37] Guacho, G. B., Abdali, S., Shah, N., Papalexakis, E. E., 2018. Semi-supervised content-based detection of misinformation via tensor embeddings. CoRR abs/1804.09088.

[38] Gupta, A., Kumaraguru, P., Castillo, C., Meier, P., 2014. Tweetcred: Real-time credibility assessment of content on twitter. In: International Conference on Social Informatics. Springer, pp. 228–243.

[39] Hamidian, S., Diab, M., 2015. Rumor detection and classification for twitter data. In: Proceedings of the 5th International Conference on Social Media Technologies, Communication, and Informatics, SOTICS, IARIA. pp. 71–77.

[40] Hardalov, M., Koychev, I., Nakov, P., 2016. In search of credible news. In: International Conference on Artificial Intelligence: Methodology, Systems, and Applications. Springer, pp. 172–180.

[41] Harshman, R., 1970. Foundations of the parafac procedure: Models and conditions for an "explanatory" multi-modal factor analysis. UCLA Working Papers in Phonetics 16.

[42] Hawkes, A. G., 1971. Spectra of some self-exciting and mutually exciting point processes. Biometrika 58 (1), 83–90.

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

[44] Hochreiter, S., Schmidhuber, J., Nov. 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

[45] Horne, B. D., Adali, S., Mar 2017. This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News. arXiv e-prints, arXiv:1703.09398.

[46] Hosseinimotlagh, S., Papalexakis, E. E., 2018. Unsupervised content-based identification of fake news articles with tensor decomposition ensembles. In: WSDM 2018 MIS2: Misinformation and Misbehavior Mining on the Web Workshop. pp. 1–8.

[47] Hu, X., Tang, J., Liu, H., 2014. Online social spammer detection. In: Proceedings of the 28th AAAI Conference on Artificial Intelligence. Vol. 14. pp. 59–65.

[48] Ito, J., Song, J., Toda, H., Koike, Y., Oyama, S., 2015. Assessment of tweet credibility with lda features. In: Proceedings of the 24th International Conference on World Wide Web. WWW '15 Companion. ACM, pp. 953–958.

[49] Jacovi, A., Sar Shalom, O., Goldberg, Y., Nov. 2018. Understanding convolutional neural networks for text classification. In: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Association for Computational Linguistics, Brussels, Belgium, pp. 56–65.

[50] Jin, Z., Cao, J., Zhang, Y., Luo, J., 2016. News verification by exploiting conflicting social viewpoints in microblogs. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI-16). pp. 2972–2978.

[51] Jin, Z., Cao, J., Zhang, Y., Zhou, J., Tian, Q., 2017. Novel visual and statistical image features for microblogs news verification. IEEE transactions on multimedia 19 (3), 598–608.

[52] Kang, C., Goldman, A., 2016. In washington pizzeria attack, fake news brought real guns. The New York Times 5.

[53] Knapp, R. H., 1944. A psychology of rumor. Public opinion quarterly 8 (1), 22–37.

[54] Kochkina, E., Liakata, M., Zubiaga, A., 2018. All-in-one: Multi-task learning for rumour verification. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 3402–3413.

[55] Kumar, K. P. K., Geethakumari, G., 2014. Detecting misinformation in online social networks using cognitive psychology. Human-centric Computing and Information Sciences 4 (1), 14.

[56] Kwon, S., Cha, M., Jung, K., 2017. Rumor detection over varying time windows. PloS one 12 (1).

[57] Kwon, S., Cha, M., Jung, K., Chen, W., Wang, Y., 2013. Prominent features of rumor propagation in online social media. In: Proceedings of the 2013 IEEE 13th International Conference on Data Mining (ICDM). IEEE, Dallas, Texas, USA, pp. 1103–1108.

[58] Lafferty, J. D., McCallum, A., Pereira, F. C. N., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. ICML '01. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 282–289.

[59] Le, Q., Mikolov, T., 2014. Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32. ICML'14. JMLR.org, pp. II–1188–II–1196.

[60] LeCun, Y., Kavukcuoglu, K., Farabet, C., 2010. Convolutional networks and applications in vision. In: Proceedings of 2010 IEEE International Symposium on Circuits and Systems. IEEE, pp. 253–256.

[61] Lee, Y., Yeh, Y., Wang, Y. F., 2013. Anomaly detection via online oversampling principal component analysis. IEEE Transactions on Knowledge and Data Engineering 25 (7), 1460–1470.

[62] Liu, X., Nourbakhsh, A., Li, Q., Fang, R., Shah, S., 2015. Real-time rumor debunking on twitter. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. ACM, pp. 1867–1870.

[63] Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., Cha, M., 2016. Detecting rumors from microblogs with recurrent neural networks. In: IJCAI'16 Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. New York, NY, USA, pp. 3818–3824.

[64] Ma, J., Gao, W., Wei, Z., Lu, Y., Wong, K.-F., 2015. Detect rumors using time series of social context information on microblogging websites. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. ACM, Melbourne, VIC, Australia, pp. 1751–1754.

[65] Ma, J., Gao, W., Wong, K.-F., 2017. Detect rumors in microblog posts using propagation structure via kernel learning. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vol. 1. pp. 708–717.

[66] Magdy, A., Wanas, N., 2010. Web-based statistical fact checking of textual documents. In: Proceedings of the 2Nd International Workshop on Search and Mining User-generated Contents. SMUC '10. ACM, pp. 103–110.

[67] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. NIPS'13. Curran Associates Inc., USA, pp. 3111–3119.

[68] Mitra, T., Gilbert, E., 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. pp. 258–267.

[69] Mitra, T., Wright, G. P., Gilbert, E., 2017. A parsimonious language model of social media credibility across disparate events. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. ACM, New York, NY, USA, pp. 126–145.

[70] Newman, N., Dutton, W. H., Blank, G., 2012. Social media in the changing ecology of news: The fourth and fifth estates in britain. International Journal of Internet Science 7 (1), 6–22.

[71] Page, L., Brin, S., Motwani, R., Winograd, T., November 1999. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab.

[72] Pérez-Rosas, V., Mihalcea, R., 2015. Experiments in open domain deception detection. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, pp. 1120–1125.

[73] Potthast, M., Köpsel, S., Stein, B., Hagen, M., 2016. Clickbait detection. In: European Conference on Information Retrieval. Springer, pp. 810–817.

[74] Procter, R., Vis, F., Voss, A., 2013. Reading the riots on twitter: methodological innovation for the analysis of big data. International journal of social research methodology 16 (3), 197–214.

[75] Qazvinian, V., Rosengren, E., Radev, D. R., Mei, Q., 2011. Rumor has it: Identifying misinformation in microblogs. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Edinburgh, Scotland, UK, pp. 1589–1599.

[76] Qin, Y., Wurzer, D., Lavrenko, V., Tang, C., 2016. Spotting rumors via novelty detection. CoRR abs/1611.06322.

[77] Quinlan, J. R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[78] Rubin, V., Conroy, N., Chen, Y., Cornwell, S., 2016. Fake news or truth? using satirical cues to detect potentially misleading news. In: Proceedings of the NAACL-CADD2016 Second Workshop on Computational Approaches to Deception Detection. San Diego, California, USA, pp. 7–17.

[79] Rubin, V. L., Chen, Y., Conroy, N. J., 2015. Deception detection for news: three types of fakes. Proceedings of the Association for Information Science and Technology 52 (1), 1–4.

[80] Rubin, V. L., Lukoianova, T., 2015. Truth and deception at the rhetorical structure level. Journal of the Association for Information Science and Technology 66 (5), 905–917.

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

[82] Santia, G. C., Williams, J. R., 2018. Buzzface: A news veracity dataset with facebook user commentary and egos. In: Proceedings of the 12th International AAAI Conference on Web and Social Media. pp. 531–540.

[83] Shi, B., Weninger, T., 2016. Fact checking in heterogeneous information networks. In: Proceedings of the 25th International Conference Companion on World Wide Web. WWW '16 Companion. International World Wide Web Conferences Steering Committee, pp. 101–102.

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

[85] Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H., 2017. Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter 19 (1), 22–36.

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

[87] Song, C., Tu, C., Yang, C., Liu, Z., Sun, M., Nov 2018. CED: Credible Early Detection of Social Media Rumors. arXiv e-prints, arXiv:1811.04175.

[88] Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., de Alfaro, L., 2017. Some like it hoax: Automated fake news detection in social networks. CoRR abs/1704.07506.

[89] Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A., 2018. Fever: a large-scale dataset for fact extraction and verification. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Association for Computational Linguistics, pp. 809–819.

[90] Tolmie, P., Procter, R., Randall, D. W., Rouncefield, M., Burger, C., Wong Sak Hoi, G., Zubiaga, A., Liakata, M., 2017. Supporting the use of user generated content in journalistic practice. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. ACM, pp. 3632–3644.

[91] Tolmie, P., Procter, R., Rouncefield, M., Liakata, M., Zubiaga, A., Jan. 2018. Microblog analysis as a program of work. ACM Transactions on Social Computing 1 (1), 1–40.

[92] Trabelsi, A., Zaiane, O. R., 2014. Mining contentious documents using an unsupervised topic model based approach. In: Proceedings of the 2014 IEEE International Conference on Data Mining. pp. 550–559.

[93] Vieweg, S., 2010. Microblogged contributions to the emergency arena: Discovery, interpretation and implications. In: Proceedings of the 2010

ACM Conference on Computer Supported Cooperative Work. ACM, pp. 241–250.

[94] Vlachos, A., Riedel, S., 2014. Fact checking: Task definition and dataset construction. In: Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science. Association for Computational Linguistics, pp. 18–22.

[95] Volkova, S., Shaffer, K., Jang, J. Y., Hodas, N., 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Vol. 2. pp. 647–653.

[96] Vosoughi, S., 2015. Automatic detection and verification of rumors on twitter. Ph.D. thesis, Massachusetts Institute of Technology.

[97] Vosoughi, S., Mohsenvand, M., Roy, D., 2017. Rumor gauge: predicting the veracity of rumors on twitter. ACM Transactions on Knowledge Discovery from Data (TKDD) 11 (4), 50.

[98] Wang, F., Lan, M., Wu, Y., 2017. Ecnu at semeval-2017 task 8: Rumour evaluation using effective features and supervised ensemble models. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 491–496.

[99] Wang, S., Moise, I., Helbing, D., Terano, T., July 2017. Early signals of trending rumor event in streaming social media. In: 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC). Vol. 2. pp. 654–659.

[100] Wang, S., Terano, T., 2015. Detecting rumor patterns in streaming social media. In: Proceedings of the 2015 IEEE International Conference on Big Data (Big Data). pp. 2709–2715.

[101] Wang, W. Y., 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, pp. 422–426.

[102] Widmer, G., Kubat, M., 1996. Learning in the presence of concept drift and hidden contexts. Machine Learning 23 (1), 69–101.

[103] Wu, K., Yang, S., Zhu, K. Q., 2015. False rumors detection on sina weibo by propagation structures. In: Proceedings of the 2015 IEEE 31st International Conference on Data Engineering (ICDE). IEEE, pp. 651–662.

[104] Wu, Y., Agarwal, P. K., Li, C., Yang, J., Yu, C., 2014. Toward computational fact-checking. The Proceedings of the VLDB Endowment (PVLDB) 7 (7), 589–600.

[105] Yang, F., Liu, Y., Yu, X., Yang, M., 2012. Automatic detection of rumor on sina weibo. In: Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics. ACM, p. 13.

[106] Yu, F., Liu, Q., Wu, S., Wang, L., Tan, T., 2017. A convolutional approach for misinformation identification. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. IJCAI'17. AAAI Press, pp. 3901–3907.

[107] Zeng, L., Starbird, K., Spiro, E. S., 2016. #unconfirmed: Classifying rumor stance in crisis-related social media messages. In: Proceedings of the 10th International AAAI Conference on Web and Social Media. pp. 747–750.

[108] Zhang, H., Fan, Z., Zheng, J.-h., Liu, Q., 2012. An improving deception detection method in computer-mediated communication. Journal of Networks 7 (11), 1811–1816.

[109] Zhao, Z., Resnick, P., Mei, Q., 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In: Proceedings of the 24th

International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, Florence, Italy, pp. 1395–1405.

[110] Zhou, L., Twitchell, D. P., Qin, T., Burgoon, J. K., Nunamaker, J. F., Jan 2003. An exploratory study into deception detection in text-based computer-mediated communication. In: 36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the. pp. 10–.

[111] Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., Procter, R., Feb. 2018. Detection and resolution of rumours in social media: A survey. ACM Comput. Surv. 51 (2), 32:1–32:36.

[112] Zubiaga, A., Kochkina, E., Liakata, M., Procter, R., Lukasik, M., Bontcheva, K., Cohn, T., Augenstein, I., 2018. Discourse-aware rumour stance classification in social media using sequential classifiers. Information Processing & Management 54, 273–290.

[113] Zubiaga, A., Liakata, M., Procter, R., Oct 2016. Learning Reporting Dynamics during Breaking News for Rumour Detection in Social Media. arXiv e-prints, arXiv:1610.07363.

[114] Zubiaga, A., Liakata, M., Procter, R., 2017. Exploiting context for rumour detection insocial media. In: Ciampaglia, G. L., Mashhadi, A., Yasseri, T. (Eds.), Social Informatics: 9th International Conference. Springer International Publishing, Cham, pp. 109–123.

[115] Zubiaga, A., Liakata, M., Procter, R., Bontcheva, K., Tolmie, P., 2015. Towards detecting rumours in social media. In: AAAI Workshop: AI for Cities. pp. 35–41.

[116] Zubiaga, A., Liakata, M., Procter, R., Wong Sak Hoi, G., Tolmie, P., 03 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. PLOS ONE 11 (3), 1–29.