Advanced Review

Credibility in social media: opinions, news, and health information—a survey

Marco Viviani* and Gabriella Pasi

In the Social Web scenario, where large amounts of User Generated Content diffuse through Social Media, the risk of running into misinformation is not negligible. For this reason, assessing and mining the credibility of both sources of information and information itself constitute nowadays a fundamental issue. Credibility, also referred as believability, is a quality perceived by individuals, who are not always able to discern with their cognitive capacities genuine information from the fake one. For this reason, in the recent years several approaches have been proposed to automatically assess credibility in Social Media. Most of them are based on data-driven models, i.e., they employ machine-learning techniques to identify misinformation, but recently also model-driven approaches are emerging, as well as graph-based approaches focusing on credibility propagation. Since multiple social applications have been developed for different aims and in different contexts, several solutions have been considered to address the issue of credibility assessment in Social Media. Three of the main tasks facing this issue and considered in this article concern: (1) the detection of opinion spam in review sites, (2) the detection of fake news and spam in microblogging, and (3) the credibility assessment of online health information. Despite the high number of interesting solutions proposed in the literature to tackle the above three tasks, some issues remain unsolved; they mainly concern both the absence of predefined benchmarks and gold standard datasets, and the difficulty of collecting and mining large amount of data, which has not yet received the attention it deserves.

How to cite this article:
WIREs Data Mining Knowl Discov 2017, 7:e1209. doi: 10.1002/widm.1209

INTRODUCTION

Nowadays, in the Web 2.0 era, the interaction between users is promoted by a number of Social Media that facilitate the establishment of social relationships, and the diffusion of information in the form of User Generated Content (UGC). In this context, users are able to publish directly textual reviews or opinions, pictures, videos, and others, almost without any form of trusted external control. It clearly emerges how, in this situation, the risk of running into misinformation is not negligible. For this reason, assessing and mining the credibility of both pieces and sources of information constitutes nowadays a fundamental issue for users. Credibility, also referred as believability, is a concept that has been studied since ancient times, and in different research fields, such as communication, psychology, and social sciences. Recent studies have identified credibility as a quality perceived by individuals, who in most cases do not have the necessary means to discern with their own cognitive capacities genuine information from fake one. With respect to the 'offline' world, where people traditionally reduce uncertainty about credibility based either on the reputation of the source of information (e.g., experts and/or opinion leaders), or on personal trust based on first-hand experiences, in the digital realm the assessment

*Correspondence to: marco.viviani@disco.unimib.it

Department of Informatics, Systems and Communication (DISCo), University of Milano-Bicocca, Milano, Italy

Conflict of interest: The authors have declared no conflicts of interest for this article.

of credibility is often more complex than in previous media contexts. Online, the multiplicity of sources involved in the dissemination of contents, their possible anonymity, the absence of standards for information quality, the ease in manipulating and altering the information, the lack of clarity of the context, and the presence of many potential targets of credibility evaluation (i.e., the content, the source, and the medium), make urgent the need of developing systems for helping users in automatically assess credibility of information and information sources.

In computer science, numerous are the approaches that have been proposed to tackle this issue. Credibility perceptions are evaluated in terms of multiple characteristics connected to the source of information, the content, and the media across which information diffuses. Most of the state-of-the-art approaches discussed in this article are based on data-driven models, i.e., they employ machine-learning techniques that classify pieces and/or sources of information as credible and not credible. Recently, also model-driven approaches based on Multicriteria Decision Making (MCDM) are emerging, which focus on aggregation schemes to assess an overall credibility estimate. Finally, graph-based approaches exploiting the social structure of connected entities have been recently investigated; they mainly focus on the concepts of credibility and/or trust propagation.

Since the Social Web is characterized by multiple social applications, which have been developed to address different aims and in different contexts, the solutions that have been proposed to address the credibility assessment issue in Social Media change depending on the considered scenario. Three of the main tasks that have tackled this issue are: (1) the detection of opinion spam in review sites, (2) the detection of fake news and spam in microblogging, and (3) the credibility assessment of online health information. In this article, we provide a review of the main and more significant approaches that have been designed to address the above tasks, by discussing their pros and cons. Despite the high number of interesting proposed solutions, some open issues still concern the problem of the absence of predefined benchmarks and gold standard datasets, and the problem of collecting and mining large amount of data, which has not yet received the attention it deserves.

The Post-Truth Era

The Oxford Dictionary has recently declared *post-truth* to be its international word of the year for 2016. It is defined by the dictionary as an adjective 'relating to or denoting circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief.' Although the concept of post-truth has been in existence for the past decade, the editors of the Oxford Dictionary have observed a spike in frequency in 2016 compared to previous years (an increase of about 2000% in the use of the term). This term has been particularly employed in the context of the EU referendum on 'Brexit' in the United Kingdom, and in the context of the presidential election in the United States.

Social Media are today seen as a long-equal and democratic instrument for exchange and sharing of information; despite this, they can turn into palaces of mirrors in which everyone seeks and finds possible confirmations to their opinions and reflections. This way, people see only themselves, their anger and discomfort, without a critical verification of the posted content. In an interesting article in the Washington Post, the *echo-chamber* phenomenon on Social Media[1] is perfectly described: there is no truth of the facts, because each has selected and receives only the news and comments with which he or she agrees in advance. In the near future it will be more and more crucial for both traditional and Social Media to address the credibility assessment issue, facing the growing importance and weight of the phenomenon formed by the swirling spread of misinformation, with the subsequent advent of the post-truth.

CREDIBILITY

In the field of communication, *credibility* is one of the oldest concepts that have been investigated. Over the years, depending on the context, credibility has been in turn associated with believability, trustworthiness, perceived reliability, expertise, accuracy, and with numerous other concepts or combinations of them.[2] In particular, the attention raised by both credibility and credibility assessment has gradually moved from traditional communication environments, characterized by interpersonal and persuasive communication, to mass communication and interactive mediated communication, with particular reference to online communication.[2,3] In this section, we first provide a brief historical background on the concept of credibility; then we illustrate its

application to computer science, with particular reference to online information. We present the open issues connected to credibility assessment in UGC, where information is generated and shared at a high rate and big volumes through Social Media.

An Overview on Credibility

One of the oldest sources involving the concept of credibility is the *Phaedrus* (around 370 BC) by Plato, a dialogue between Socrates, the protagonist, and Phaedrus, his interlocutor. In this work, Plato tackled one aspect of the concept of credibility of the information source, i.e., the speaker's knowledge of truth (or at least the resemblance of truth offered by the speaker to the audience). Later on, Aristotle in his *Rhetoric* (fourth century BC) identified another group of source credibility characteristics, connected to the *ethos*—Greek term for 'character'—of the communicator. Coming to the modern era, since the 1950s researchers in the fields of psychology and communication have been interested in defining and studying source credibility. The research undertaken by Hovland and colleagues[4–6] constitutes the first systematic work about credibility and mass media. The authors studied in particular the credibility of mass communication messages, by investigating how people evaluated these messages coming from 'high credibility' and 'low credibility' sources. From the results of their studies it emerged that, even if the audience was more influenced by 'high credibility' sources, the information was equally learned from both source types. In fact, as illustrated by Fogg and Tseng,[7] credibility is a *perceived* quality of the *information receiver*, and it is composed of multiple dimensions. In this sense, the process of assessing *perceived credibility* involves different *characteristics*, which can be connected to: (1) the *source* of the message, (2) the *message* itself, i.e., its structure and its content, and (3) the *media* used to diffuse the message.[8]

Most researchers agree that there are at least two key dimensions connected to (perceived) credibility: *expertise* and *trustworthiness*.[6,9] With particular reference to source credibility, expertise is 'the perceived knowledge, skill, and experience of the source.'[7] It allows people to measure to which extent a communicator is capable of making correct statements. Trustworthiness refers to which degree an audience perceives the statement made by a communicator to be valid.[6] This aspect is strictly connected to message credibility, which allows to measure, according to Liu,[10] to what extent users think information (i.e., the message) is truthful, unbiased, accurate, reputable, competent, and current. In other words, a source is trustworthy for being honest, careful in choice of words, and disinclined to deceive.[11] Information is trustworthy when it appears to be reliable, unbiased, and fair.[12] When considering these characteristics, the impact of the delivery medium can change the perception that people have about sources of information and information itself. Sometimes, depending on the context, the medium diffusing information can be misleadingly interpreted as the source of information. Thus, one important question to be tackled nowadays is whether new media introduce new factors into credibility assessment.[13] In particular, according to Metzger and Flanagin,[14] a great attention must be paid in conceptualizing credibility in the digital realm. This issue will be discussed in detail in the next section.

From Offline to Online Credibility

As previously illustrated, the problem of assessing information credibility has a long history and involves different research fields; therefore, it does not refer only to online accessible information. In the 'offline' world, long before the birth and spread of the Web, users have always been confronted with the problem of trusting information obtained via different kinds of media. Although it has been shown that the cognitive skills that people need to assess information credibility are not dependent on the scenario (i.e., offline or online),[15] these skills often depend on the way in which people get this information. In real life, people usually deal with sources of information that can be (1) *organization-oriented*, i.e., provided by well known organizations such as reputed newspapers or popular enterprises; (2) *independent*, i.e., provided by third parties such as no profit organizations or individuals considered as experts in a given field; (3) *interpersonal*, i.e., based on direct communication and knowledge among individuals. In these three cases, a common practice by which people have traditionally reduced uncertainty about credibility include judgment based on both the *reputation* of the source of information or of traditional information intermediaries such as experts and/or opinion leaders—cases (1) and (2), and the personal *trust* based on first-hand experiences with the information providers—case (3).

Since nowadays a huge amount of information is distributed online, and the Web represents for many people the primary means to stay informed, several of these traditional intermediaries have been removed through a process of 'disintermediation.'[16] In this way, according to Flanagin and Metzger,[15]

digital media 'have in many ways shifted the burden of information evaluation from professional gatekeepers to individual information consumers.' Thus, the assessment of credibility in the online environment is often more complex with respect to previous media contexts, for a number of different reasons. A first issue is connected to 'the multiplicity of sources embedded in the numerous layers of online dissemination of content,'[17] and to the fact that these sources and connected information messages become confused in information seekers' minds almost immediately after having performed online search.[18] Other issues concern the absence of standards for information quality, the ease in manipulating and altering the information, the lack of clarity of the context, and the presence of many potential targets of credibility evaluation interacting together in users' perceptions, i.e., the content, the source, the medium.[14]

Credibility and Social Media

The previously described issues to assess credibility are exacerbated in the Web 2.0 context, where users are encouraged to interact and collaborate with each other through Social Media, by playing the role of creators of UGC in a virtual community.[19] In this context, personal knowledge is replaced by virtual relationships, through which it is easier and inexpensive for users to directly exchange information. Many studies in behavioral psychology have shown that people have an innate tendency to share information in virtual communities, whether or not explicit benefits are involved.[20–22] In such a scenario, the process of evaluating the most credible information differs both from traditional media and Web 1.0 technologies.

Evaluating information credibility in the Social Web deals with the analysis of both UGC and their authors' characteristics,[23] and to the intrinsic nature of Social Media platforms.[24] Generally speaking, this means to take into account characteristics of credibility both connected to information (i.e., UGC) and to information sources, as well as to the social relationships connecting the involved entities. In this context, the users' credibility perceptions are usually based on crowd consensus, which is typical in interacting communities.[25,26]

Even if credibility is a characteristic perceived by individuals, credibility assessments should not be up to users, especially in the online environment.[3] In fact, humans have only limited cognitive capacities to effectively evaluate the information they receive, especially in situations where the complexity of the features to be taken into account increases.[27] For this reason, there is nowadays the need of developing interfaces, tools or systems that are designed to help users in automatically or semi-automatically assess information credibility. In the next section, we illustrate the approaches that have been proposed so far to assess credible information in the Social Web.

APPROACHES TO CREDIBILITY ASSESSMENT

Social Media refers to a plethora of different Web 2.0 Internet-based applications, which take the form of blogs, social networks, forums, microblogs, photo sharing, video sharing, review sites, and so on. Depending on both the context and the aim to which a given social application is developed, several different characteristics can concur to the assessment of information credibility. In this literature survey, we review in particular the main approaches that have been proposed for: (1) opinion spam detection, with particular reference to fake review detection in review sites, (2) fake news and spam detection in microblogging sites, and (3) credibility assessment of online health information, in particular in healthcare social forums.

Opinion Spam Detection in Review Sites

The first studies on *opinion spam detection* were proposed by Jindal and colleagues in the late 2000s,[28–30] in conjunction with the spread of both opinion-sharing sites and the first social platforms allowing users to diffuse their opinions in the form of reviews concerning given entities, e.g., products, restaurants, services, and so forth. *Opinion spam* covers malicious activities attempting to mislead either human readers, or automated opinion mining and sentiment analysis systems. These fraudulent activities concern both the formulation of undeserving positive opinions to promote some target entities, and the provision of false-negative opinions to some other entities, in order to damage their reputation. In this context, opinion spam can have different forms; in most cases, it concerns *fake reviews*, but it can also be in the form of fake comments, fake blog and social network postings, and deceptive messages.

This section focuses in particular on *fake review detection* in *review sites* (e.g., TripAdvisor, Yelp, Amazon, etc.); in last years, this task has received a growing attention within the scientific community, and several approaches have been proposed to date. In fact, online reviews represent one of the consumer's primary factors in choosing a product/service, and there is an increasing interest of

companies in customers' online opinions. This can, in fact, help companies to reshape their business by adapting products, services, and marketing to consumer preferences.[31] For this reason, fake reviews can damage both the buyer, whose choice is guided to nonoptimal decisions, and the business, which not only looses consumer's trust, but also risks to misinterpret users' needs.

A well-known spam review categorization has been provided by Jindal and Liu,[30] according to whom there are generally three main types of fake reviews:

- *Untruthful reviews*: deliberately providing undeserving positive or negative reviews to some entities in order to promote or to scuttle them, misleading readers or opinion mining systems;
- *Reviews on brands*: considered as spam because they focus specifically on the brands, the manufacturers or the sellers of the products, by disregarding the specific products or services;
- *Nonreviews*: occurring in the form of advertisements or other kinds of posts that do not contain opinions (e.g., questions, answers, and random texts).

The majority of the approaches proposed in the literature for fake review detection consider the first category of reviews (i.e., untruthful reviews), for two main reasons[32]: (1) they undermine the integrity of the online review system or site in exam; (2) it is almost impossible to distinguish between fake and genuine reviews by only reading their content. Therefore, to build automatic or semi-automatic methods for the recognition of untruthful reviews, both *data-driven* and *model-driven* approaches have been proposed and developed, based on data mining and machine-learning techniques, and MCDM, respectively. Besides the approaches that directly identify fake reviews, some other approaches have been proposed to the aim of capturing *spammers* and *group spammers* who spread untruthful reviews in opinion-sharing platforms.

Whatever the purpose, to directly identify false reviews or users that spread them, both strategies usually consider (simultaneously or not) a certain number of characteristics, namely *features*, that can be evaluated in terms of credibility. They can be extracted from both (1) the *review* and (2) the *reviewer*. In the former case, it is possible to deal with *linguistic* features, associated with the content of the review, and *meta-data* features, i.e., additional information connected to the textual information. In the second case, we can deal with characteristics connected to the behavior of the reviewer, i.e., *behavioral* features, and with features connected to the user's profile (if available). In addition, even if they have been less used in the literature so far, other features connected to (3) the *network structure* represented by the relationships among users or other entities can be taken into account; they can be defined as *social* features. Finally, another category of less used characteristics includes *product* features, directly linked to (4) the *product/service* reviewed. While fake review detection approaches are based in particular on linguistic, behavioral, meta-data and product features, behavioral, and social features are nowadays used in most cases to detect spammers and group spammers.

Two literature surveys on opinion spam detection in review sites have been recently published.[31,32] Aim of this section is to describe, among the approaches that have been well presented and summarized in the above surveys, the ones that have demonstrated to be the more effective. In addition, we provide an overview of the most recent approaches, and with respect to previous surveys we organize the presentation of the state of the art by focusing on the aim for which the approaches analyzed have been proposed: (1) fake review detection, and (2) spammers/group spammers detection. Within each of the above classes of approaches, we focus on the features that they employ.

Fake Review Detection

As previously outlined, fake review detection concerns, in most cases, the identification of untruthful reviews within an opinion-sharing system. Based on the features on which the proposed approaches focus, we categorize them into purely *content-based* and *multi-feature-based* approaches. Content-based approaches are based on linguistic features, while multi-feature-based approaches employ multiple kinds of features among those previously described. The following subsections summarize the approaches in each of the above categories.

Content-Based Approaches

Several approaches proposed in the literature are based exclusively on features extracted from the text of reviews. Different linguistic features can be considered; the simplest ones are either individual words or *n*-grams, (i.e., groups of *n* words), which can be either weighted or unweighted.[30,33] For a complete overview on the multiple kinds of linguistic features that have been proposed in the literature, refer to Crawford et al.[32] The majority of purely content-

based approaches are based on supervised learning techniques.[34–37]

Ott et al.[34] use three groups of features to train Nave Bayes (NB) and support vector machine (SVM) classification algorithms. For each review, the authors define linguistic features that are based on the frequencies of each parts-of-speech (POS) tag,[38] and they test differences in POS distribution between fake and nonfake reviews based on the genre of reviews. In addition, they consider psycholinguistic deception detection by using the Linguistic Inquiry and Word Count (LIWC) software,[39] which assigns 80 psycholinguistic meanings to 4500 keywords. Finally, they consider 'classical' *n*-gram feature sets, i.e., unigrams, bigrams, and trigrams. The evaluation of this approach is based on a gold standard dataset of genuine reviews collected from review sites and fake reviews generated by means of the Amazon Mechanical Turk (AMT). With respect to considering only simple *n*-gram features, this approach achieves better results, even if it does not outperform models considering multiple features of different nature (i.e., in addition to linguistic ones).[30] In a later work,[35] the same authors focus on the optimization of the supervised classification method, obtaining slightly better result with respect to their previous approach. On the same dataset, Banerjee and Chua[36] use logistic regression (LR) to build a classification model by considering the readability of a review (i.e., its complexity and reading difficulty), the distribution of POS tags, and the review writing style, i.e., positive cues, perceptual words, and usage of future tense. Fusilier et al.[37] use both character *n*-grams and word *n*-grams obtaining the best results with character *n*-grams with values for *n* of 4 and 5, respectively, by using NB and an SVM classifier on the Ott et al.'s dataset.[34]

It is worth to be noticed that Mukherjee et al.[40,41] and Sun et al.[42] observed that using the AMT for producing fake reviews causes inaccurate results in review spam detection, because fake reviews generated ad hoc (e.g., by means of AMT) are not meaningful in spam detection techniques when compared with real reviews collected from opinion-sharing sites. In other words, the employed dataset is biased because the motivations and the psychological states of mind of people involved in the AMT crowdsourcing activity can be quite different from those of the professional spammers in the real world. This explains the good results obtained by previous approaches in fake review detection based only on linguistic features.

Other works propose the use of *language models*[43] to represent the textual content of reviews. Lai et al.[44] employ this content-based representation to ascertain possible 'concept substitutions' in untruthful reviews: if the semantic content of a review is mainly generated from another review, this suggests that both reviews may not truly reflect writers' opinions. To evaluate the likelihood that a review has been generated by another one, the Kullback–Leibler divergence[45] is employed. To convert this measure to a spam score for each review, the authors us a linear normalization method. Finally, they employ an SVM-base method to classify spam and genuine reviews, by obtaining a declared 81% precision (it should be noticed that the precision of the Jindal and Liu method,[30] which employs multiple kinds of features, is 85%). Lau et al.[46] extend the previous work by proposing a semantic language model (SLM)-based approach. By using both the semantic relationships captured from WordNet, and high-order concept associations discovered via the proposed text mining method, the SLM method detects duplicate untruthful reviews even if their wordings are not the same. The proposed model achieves a declared true positive rate of over 95% in fake review detection. Both approaches previously described[44,46] employ as a gold standard a real-world dataset collected from Amazon, and manually labeled by human experts.

Generating gold standard datasets for evaluations is a complex task in the fake review detection context.[47] In fact, as it emerges from the presentation of the approaches illustrated so far, current methods of collecting reviews labeled with respect to credibility are neither effective (AMT) nor practical (when human intervention is required). Therefore, in the absence of large amounts of labeled examples of deceptive and truthful opinions, some approaches have proposed the use of semi-supervised learning techniques, by considering only a few examples of deceptive opinions and a set of unlabeled data. Specifically, these methods are based on the so-called *positive-unlabeled* (PU) *learning*.[48] Hereafter, we describe those PU-based methods acting on linguistic features. Ren et al.'s research work on fake review detection[49] assume the presence of some differences in the language structure and sentiment polarity between deceptive reviews and truthful ones.[50] The authors identify a set of features related to the text of the review and combine two clustering algorithms to identify deceptive reviews. PU learning is implemented in a subsequent approach: MPIPUL,[49] which identifies deceptive reviews based on both Latent Dirichlet Allocation (LDA)[51] and SVM. Experiments are performed on a gold standard dataset obtained from the ATM dataset by Ott et al.,[34] but the

approach is only evaluated against other PU algorithms, and not w.r.t. state-of-the-art fake review detection approaches. A similar approach has been recently proposed by Fusilier et al.[52]; with respect to the original PU learning algorithm proposed by Liu et al.[53] to classify text documents, this approach is more conservative when selecting the negative examples (i.e., not deceptive opinions) from the unlabeled ones.

Multiple-Feature-Based Approaches

In the literature, the use of multiple features of different nature has proven to be a more effective method than those based on a unique feature type (i.e., linguistic features). In fact, experienced review spammers write so good fake reviews to make them indistinguishable from truthful reviews, even when they are read by experts in review spam detection.[31,54]

In the first research works on opinion spam detection,[29,30] this aspect was already clear. In their 2008 work,[30] Jindal and Liu proposed an approach to detect duplicate and near-duplicate reviews, by observing that fake reviews are often duplicates of a small number of reviews created by spammers, which are used as templates to generate false information on different products or services. To estimate the similarity between reviews and to obtain duplication scores, the authors propose to use bigrams and the Jaccard Index. By considering heterogeneous features (i.e., connected to the review and its content, to the reviewer, and to the reviewed product), this approach applies LR to detect spam reviews on a manually labeled dataset composed of 470 fake reviews from Amazon, identified by means of their computed duplication scores. This method has been evaluated by means of the area under ROC curve (AUC), a standard measure used in machine learning for assessing the model quality, obtaining a 78% AUC value by considering hybrid features with respect to the 63% AUC value obtained by using only textual features. A study by Mukherjee et al.[41] found that when considering abnormal behavioral features of reviewers, the obtained results are better than those obtained by using only the linguistic features of the reviews. The authors consider for example in their approach the maximum number of reviews written by a customer, the amount of positive reviews with respect to negative ones, the review length, the reviewer deviation in assigning ratings with respect to the general rating consensus, and the content similarity. Supervised learning approaches for deceptive review detection are defined and evaluated, by using both the AMT (AMT) fake reviews dataset (described by Ott et al.[34]), and a real-world review dataset (where reviews are automatically classified and labeled by Yelp in *recommended* and-*not recommended*). The obtained results are similar to those reported in previous studies: using only *n*-gram features performs well on the AMT dataset, but when used on a real world dataset (i.e., Yelp in this case) the approach performs significantly worse. Once again, this proves the intrinsic bias in the ground truth constructed by means of the AMT.

Li et al.[55] propose an approach constituted by two phases. First, the authors investigate the impact that review- and reviewer-based features have on fake review detection. The supervised NB method employed for this task achieves, as expected, significant improvement compared with heuristic methods; it has been tested on a small manually labeled dataset from Epinions. To boost the performance of the approach, by considering a larger number of unlabeled data, the authors employ a semi-supervised cotraining algorithm, which demonstrated its effectiveness with respect to supervised NB. Another semi-supervised approach for fake review detection has been proposed by Li et al.[56]; in their model, the authors take into consideration a subset of the features proposed by the previously described approaches, and implement a PU learning algorithm to classify Chinese reviews. The review dataset is extracted from the review hosting site dianping.com, which is the Chinese equivalent of Yelp. Even if it does not propose any automatic approach for opinion spam detection in review sites, the work by Luca and Zervas[57] is interesting in the sense that it analyzes multiple kinds of features as well as the motivations (often based on economic incentives) connected to the process of issuing fake reviews. The awareness of the importance of features in terms of credibility can lead to the development of model-driven approaches instead of data-driven ones.

Recently, Viviani and Pasi[58] have proposed a MCDM approach for fake review detection, where multiple kinds of features, both review- and reviewer-based are considered and aggregated by suitable functions to obtain an overall credibility estimate. This method is based in particular on quantifier guided aggregations of credibility scores associated with the considered features: at least k or at least k *important* features must characterize a review to classify it as true rather than fake. The labeled dataset used for evaluation purposes has been crawled from Yelp, as in the work by Mukherjee et al.[41] The outcome of this work is that an MCDM-based approach can achieve the same results obtained with the best machine-learning classifiers.

Spammer Detection

With respect to fake review detection, the approaches that have been proposed for spammer and group spammers detection are more recent. In most cases, they focus on behavioral and social features, and nowadays they are increasingly exploiting the graph-based structure that represents the relationships among various entities (e.g., reviews, reviewers, stores, and products).

Behavioral-Based Approaches

The approach by Lim et al.[59] involves the use of behavioral heuristics to detect spammers, for example their attitude to target specific products or product groups in order to maximize their impact, and the fact that spammers deviate from the other reviewers in their ratings of products. Considering these aspects, the authors propose a scoring method for ranking reviewers by implementing some measures that capture spamming behaviors on an Amazon review dataset. A subset of highly suspicious reviewers is then identified by human experts, and a regression model is employed to automatically score reviewers. The authors found that deleting reviews written by suspicious reviewers has a high impact on the rating of the target products. As pointed out by Heydari et al.,[31] one of the possible issues of this approach is that, by considering the rating deviation of a reviewer with respect to the product average rank as an evidence of anomaly, there is the risk to ignore that this could simply represent a different personal experience of the reviewer with respect to others. Fei et al.[60] propose the idea of measuring the 'burstiness' that characterizes reviews generated by spammers to identify fake reviews. Bursts are reviews that become very popular and receive great attention in a given time period and/or in a certain area. Due to their rapid diffusion, they become suspicious as well as their authors. A kernel density estimation (KDE)[61] technique is used to detect review bursts, and both behavioral features for spammers and features extracted from review bursts are considered. To identify in an automatic way spam reviewers, the authors employ a Markov random field (MRF)[62,63] model coupled with a loopy belief propagation (LBP)[64] algorithm, on the Jindal et al.'s[30] dataset.

Mukherjee et al.[65] consider the particular case of spammers who act with other spammers to achieve a given goal. They present Group Spam Rank (*GSRank*), an approach for group spammers detection, which exploits some behavioral features connected to groups of spammers (e.g., content similarity between members, content similarity among a group, group size, etc.). By considering these features, it is possible to provide an effective classification of reviewers into spam groups and nonspam groups. In this work, multiple experts were employed to create a labeled group opinion spammer dataset, consisting of 2412 labeled spam and nonspam groups.

Graph-Based Approaches

These approaches exploit the graph-based nature of the relationships that can occur between entities to be evaluated in opinion spam/spammer detection. The first graph-based methods to detect review spammers were proposed by Wang et al.[66,67] These approaches consider three entities as nodes of the graph: the review, the reviewer and the store (i.e., the reviewed entity). The authors explore how the interactions between nodes in the graph can be used to identify spam, and propose an iterative computation model to identify suspicious reviewers. They obtain data for experiments by crawling resellerratings.com, one of the largest hosts of store reviews. First, they let the proposed algorithms identify highly suspicious spammer candidates; then, human judges examine these candidates to decide how many reviewers are indeed suspicious. The evaluation of this approach produces both good precision and good human evaluators agreement.

The approach described by Akoglu et al.[68] proposes a framework, *FraudEagle*, for detecting spammers and fake reviews in online review sites. This framework exploits the connectivity structure linking users, products, and reviews, by observing that spammers are usually linked to good/bad products with negative/positive fake reviews, and vice versa. For this reason, it is possible to 'sign' with *sentiment* the network edges, and to apply a signed inference algorithm extending LBP to infer the class labels of users, products, and reviews. However, with respect to prior work,[67] it does not improve the spam/spammer detection accuracy. In a recent work, Ye and Akoglu[69] propose *GroupStrainer*, a two-step method to discover group spammers and their targeted products. The considered review network is a bipartite graph consisting of reviewer nodes connected to product nodes through review relations. The authors define a Network Footprint Score (NFS) measure to quantify the likelihood of products of being spam targets. Then, a clustering phase identify highly similar spammers (i.e., group spammers) on a 2-hop subgraph induced by top ranked products (i.e., products having the highest NFS values). In the experiments, the authors compare the proposed method to various graph anomaly detection methods on synthetic

datasets, and they study its performance on two large real-world datasets (Amazon and iTunes).

Other recent approaches focusing on group spammer detection are proposed by Choo et al.[70] and Wang et al.[71] In the former approach, the authors explore the community structures to distinguish spam communities from nonspam ones with sentiment analysis on user interactions. Through experiments over a crawled Amazon dataset, the authors obtain comparable results with respect to state-of-the-art approaches, and they find—not surprisingly—that strong positive communities are more likely to be opinion group spammers. The latter approach illustrates the 'loose' group spam problem, i.e., each group member is not required to review every target product. This problem is addressed by using bipartite graph projection. The authors propose a set of group spam indicators to measure the 'spamicity' of a loose group of spammers, and they design an algorithm to identify highly suspicious loose group spammers in a divide and conquer manner. Experimental results show that the proposed approach achieves both target goals, i.e., the detection of loose group spammers, and the detection of group spammers, with better results with respect to frequent itemset mining (FIM)[72] approaches.

Discussion

Several approaches presented in this section have been defined in the last years to tackle the problem of opinion spam detection in review sites. As previously illustrated, most of the proposed approaches are based on supervised or semi-supervised techniques that make use on multiple kinds of characteristics that can be related to credibility. One of the main open issues that these approaches do not solve is the lack of a valuable gold standard to train and evaluate the proposed classifiers; in fact, in most cases, pseudo fake reviews have been employed rather than real fake ones.[73] Some approaches consider duplicate reviews as fakes, not considering that similar review contents might be produced by honest reviewers; moreover, spammers commonly copy the text of truthful reviews to generate fakes.[31] The human intervention for labeling reviews can be inapplicable to big amount of data, and it is debatable as also the reliability of the human assessors may be questioned. Even the usage of crowdsourcing platforms such as the AMT for producing fake reviews has been proved unreliable, due to the fact that the motivations and the psychological states of mind of people using the AMT can be quite different from those of professional spammers in the real world.[40,42] Recent approaches have proposed to use as a gold standard a subset of both *recommended* and *non-recommended* reviews automatically labeled by some review sites, namely Yelp and Dianping.[41,56,57] This constitutes a quite strong assumption, since these review sites do not disclose their policy for detecting nonrecommended (fake) reviews, and even in such a case the filtered reviews would be affected by some uncertainty related to the applied decision process. It is clear that none of the solutions proposed so far can be considered fully reliable, and that new approaches to the evaluation of credibility are necessary. To tackle the problem of opinion spam detection from a different perspective, recent approaches focus on identifying spammers and group spammers in Social Media, by using both behavioral characteristics of reviewers and the social graph connecting multiple entities (i.e., users, products, and reviews) in review sites. Table 1 lists the approaches to fake review and spammer detection, by summarizing the main characteristics of each approach.

Fake News Detection in Microblogging

Microblogging is a well-established broadcast medium that allow users 'to exchange small elements of content such as short sentences, individual images, or video links.'[74] These short messages are commonly referred as microblog posts, microposts, or status updates. Nowadays, thanks to the success of the microblogging platform Twitter,[a] they are also simply referred as tweets. In Social Media classification, microblogs stand in fact between traditional blogs and social networking sites, and they are characterized by both a high degree of self-presentation, and a medium to low degree of social presence and media richness.[75]

This preamble on microblogging is necessary to underline that, due to its sharing characteristics and to the fact that messages display a wide variety of content, the credibility assessment of information on this kind of medium may concern different aspects. As outlined in the literature,[76] short messages in microblogging can correspond both to the so called *conversation items* (e.g., chats, personal updates, gossips, etc.), which concern the user and its circle of friends, and to *news items*, representing more general and relevant information.[77–79] This latter kind of information can influence a broader community, in particular when microblogs deal with the well known *trending topics*, which can be considered 'headline news or persistent news.'[80] In this context, Rubin et al.[81] identify three types of fake news: exposed fraudulent journalistic writing, humorous fakes, and large-scale hoaxes. This latter category is the one that

TABLE 1 | Summarization of the Proposed Approaches for Opinion Spam Detection in Review Sites

Authors	Year	Focus	Approach
Jindal and Liu[30]	2008	Fake review detection	Multifeature-based
Lai et al.[44]	2010	Fake review detection	Content-based
Lim et al.[59]	2010	Spammer detection	Behavioral-based
Li et al.[55]	2011	Fake review detection	Multifeature-based
Ott et al.[34]	2011	Fake review detection	Content-based
Wang et al.[66]	2011	Fake review/Spammer detection	Graph-based
Lau et al.[46]	2012	Fake review detection	Content-based
Mukherjee et al.[65]	2012	Spammer/Group spammer detection	Behavioral-based
Wang et al.[67]	2012	Fake review/Spammer detection	Graph-based
Akoglu et al.[68]	2013	Fake review/Spammer detection	Graph-based
Fei et al.[60]	2013	Fake review/Spammer detection	Behavioral-based
Mukherjee et al.[41]	2013	Fake review detection	Multifeature-based
Banerjee and Chua[36]	2014	Fake review detection	Content-based
Li et al.[56]	2014	Fake review detection	Multifeature-based
Ren et al.[49]	2014	Fake review detection	Content-based
Choo et al.[70]	2015	Group spammer detection	Graph-based
Fusilier et al.[37]	2015	Fake review detection	Content-based
Fusilier et al.[52]	2015	Fake review detection	Content-based
Wang et al.[71]	2015	Group spammer detection	Graph-based
Ye and Akoglu[69]	2015	Group spammer detection	Graph-based
Lucas and Zervas[57]	2016	Feature analysis	Content-based
Viviani and Pasi[58]	2016	Fake review detection	Multifeature-based

is better suited to be spread on Social Media like Twitter, and the one that has a higher impact on society due to the speed with which hoaxes diffuses in microblogging.

In recent years, a number of different approaches have been proposed for assessing credibility in microblogging, by also considering its time-sensitive nature. In the literature, these methods can be broadly classified into two categories[82,83]: *classification-based* approaches, and *propagation-based* approaches, which exploit the network structure of users and tweets. In addition, a number of *survey-based studies* have been presented, as it will be discussed later in this article.

Classification-Based Approaches

Castillo et al. were among the first to tackle the problem of information credibility on microblogging sites, Twitter in particular,[79,84] in a structured way, by using classification-based approaches. In their first work,[79] the authors focus on automatic methods for assessing the credibility of a given 'time-sensitive' set of tweets, i.e., a trending topic. Specifically, they analyze microblog posts (messages) related to trending topics, and they classify these topics as credible or not credible based on multiple features extracted from them. In particular, the authors identify *message-based features*, focusing on the content of the message (e.g., length, syntax, sentiment, etc.), *user-based features*, connected to the user who posted the tweet (e.g., number of friends and followers), *topic-based features*, obtained by aggregating the previous ones (e.g., fraction of positive or negative sentiment), *propagation-based features*, related to the propagation tree built by considering retweets (e.g., depth of the retweet tree). This approach is based on supervised learning, and in order to assess its effectiveness, the authors have built a suitable dataset by collecting 2500 trending topics via *TwitterMonitor*,[85] keeping only those cases containing at most 10,000 tweets each. The labeling of the dataset was carried out in two phases: in a first phase some evaluators (by means of the AMT) were requested to assess whether the collected tweets concerned news about specific events or rather they were 'conversation' topics. Then, for the subset of tweets identified as news, a second group of evaluators were requested to label them as credible or not credible. By performing

a threefold validation strategy and by applying different classifiers (i.e., SVM, decision trees, decision rules, and Bayesian networks), the authors obtained good results in identifying 'newsworthy' topics, while using a J48 decision tree[86] they reached a 86% accuracy in credibility classification with respect to a random predictor. In this paper, a feature analysis is also provided, to illustrate the contribution that different types of features give in terms of credibility assessment. The outcomes of this work are: (1) - message- and user-based features are not enough to effectively classify credible trending topics; (2) in general credible news are propagated by authors who previously wrote a considerable amount of messages; (3) propagation-based features (having many reposts) are particularly effective. Moreover, the authors outline that tweets not including URLs are in most cases related to noncredible news, while tweets including negative sentiment are related to credible content.

In another paper, Castillo et al.[84] discuss more broadly the problem of information credibility in microblogging services. By presenting a case study about information propagation and news event credibility in Twitter, first they summarize their prior work,[79] and then they redesign the learning scheme. A first supervised classifier is used to decide if an information cascade corresponds to a 'newsworthy' event. A second supervised classifier is employed to decide if this news event has to be considered as credible or not. Both classifiers are trained over labeled data, obtained using crowdsourcing tools. By introducing a label named 'unsure' to identify tweets that do not belong neither to news nor to chat messages, these 'unsure' tweets are removed from the training dataset, thus improving the performance of the approach.

Kang et al.[87] propose an approach for recommending credible information connected to specific topics in Twitter. The authors define three models that focus on different credibility indicators: (1) features extracted from the content of tweets related to specific topics (e.g., length of tweet, number of URLs, number of mentions, etc.), (2) features connected to the network-structure connecting users and tweets (e.g., 'follow' relationships, retweets, etc.), and (3) features from both the previous approaches that are combined in different ways to predict credible information and credible information sources. Each model is evaluated by using a J48 tree-based learning algorithm which is trained on a set of tweets that are manually annotated with credibility scores. Differently from Castillo et al.,[79] who propose an algorithm that acts on groups of 'newsworthy' tweets, the approach described in this paper aims at classifying each tweet individually, and at predicting user credibility. Not surprisingly, as already demonstrated in the previously described work,[79] the features that consider the underlying network and the dynamics connected to information flows are better indicators of credibility in microblogs than linguistic features.

Another work that assesses the credibility of individual tweets is the one by Gupta and Kumaraguru.[88] In this work, the authors use the SVM rank algorithm[89] (a variant of the SVM algorithm that is used to solve certain ranking problems via learning to rank[90]), to illustrate that ranking tweets based on Twitter features (i.e., content- and user-based) can help in assessing the credibility of tweets about an event. To this purpose, tweets are labeled by human annotators. To evaluate the effectiveness of the proposed approach, the standard normalized discounted cumulative gain (NDCG) metric[91] is employed. In a subsequent work, Gupta et al.[92] extend the previous work and propose *TweetCred*, a real-time system in the form of a Chrome extension. *TweetCred* takes a direct stream of tweets as input, and it computes the credibility for each tweet on a scale of 1 (low credibility) to 7 (high credibility). This research work is based on a semi-supervised ranking model using SVM Rank for assessing credibility, based on tweets related to six high impact crisis events of 2013. With respect to prior work, in this paper a more exhaustive and comprehensive set of features is used: 45 features categorized as *tweet meta-data features*, *tweet content features*, *user-based features*, *network features*, *linguistic features*, and *external resource features*. The ground truth was obtained through crowdsourcing: human assessors labeled around 500 tweets randomly selected per event. This system provides useful insights on how credibility evaluation models evolve over time. As in the case of opinion spam detection, recent approaches are focusing on the identification of spammers also in the context of fake news detection. Galán-Garca et al.[93] focus on the detection of troll profiles in Twitter. Their assumption is that, since 'every trolling profile is followed by the real profile of the user behind the trolling one […] it is possible to link a trolling account to the corresponding real profile of the user behind the fake account, analyzing different features present in the profile, connections' data and tweets' characteristics, including text, using machine-learning algorithms.' Using manually-selected genuine profiles, and collecting at least 100 genuine tweets per profile, the authors compared the performance of different classification algorithms (i.e., Random Forests, J48, *k*-Nearest-Neighbor, Sequential Minimal Optimization, and NB) on the selected features: the content of

the tweet published by the user, the time of publication, the language and geolocation, and the Twitter client. Even if this approach reached 'only' 68.47% of accuracy in its best result, this work constitutes one of the first significant approaches that try to directly detect trolls in Social Media and microblogs in particular, and it has been applied to a real case study, for the identification of the responsible of cyberbullying in a school in the city of Bilbao (Spain). Also Gupta and Kaushal[94] propose an integrated approach for spammer detection, which combines three learning algorithms, i.e., NB, Clustering, and Decision Trees, with the aim of improving spammer detection accuracy. The considered features are followers/followees, URLs, spam words, replies, and hashtags. The algorithms accuracy in the detection of non-spammers is about 99.1%, but the accuracy of detecting spammers reaches the 68.4%, which is the same result obtained by Galán-Garca et al.[93]

Propagation-Based Approaches

The approaches that focus on the concept of propagation for assessing the credibility of tweets/news, usually consider the propagation of rumors (false claims) in microblogs, by exploiting the network structure constituted by retweets, and the social graph constituted by followers and followees. Furthermore, also trust propagation on the social graph can be assessed. Mendoza et al.[95] explore the behavior of Twitter users under an emergency situation (the 2010 earthquake in Chile); a preliminary study on the dissemination of false rumors and confirmed news is reported. On a crawled dataset containing tweets and other user-related information, the approach analyzes the characteristics of the social network of the community surrounding the topic, and how trending topics propagate. The considered characteristics are the relation between the number of followers/followees, the number of tweets each user posts, and the retweet activity during the first hours of the emergency situation. From this analysis, it emerges that the network topology characteristics remain unchanged with respect to normal circumstances, and that the vocabulary used in critical situations exhibits a low variance. Concerning credibility, and the spread of rumors through the network, the authors have manually selected some *confirmed truths*, i.e., reliable news items confirmed by reliable sources, and some *false rumors*, i.e., baseless rumors that emerged during the crisis. The outcome of this credibility study is that rumors tend to be questioned much more than confirmed news by the Twitter community; therefore, the authors suggest that microblogs could implement methods to warn people by automatically reporting them highly questioned information.

With respect to Mendoza et al.,[95] the approach proposed by Seo et al.[96] studies how to identify sources of rumors when there is a limited view on the rumor provenance, and how to determine whether a piece of information is a rumor or not. The method is based on the assumption that rumors are initiated from only a small number of sources, whereas truthful information can be observed and originated by a large number of unrelated individuals. This approach relies on the use of *network monitors*, i.e., individuals who have heard a particular piece of information (from their social neighborhood), but who do not want disclose neither who told it to them, nor when they learned it. If a monitor receives a rumor, it is called positive monitor; negative monitor otherwise. To find a rumor source, the authors base their approach on the intuition that the source must be close to positive monitors and far from negative ones. For this reason, the authors introduce four metrics: the number of reachable positive monitors, the sum of distances to reachable positive monitors, the number of reachable negative monitors, and the sum of distances to reachable negative monitors. By computing these metrics for each node, it is possible to sort all the nodes in the network; in the resulting sorted list, the first node is the top suspect source of the rumor. In addition, to identify if a piece of information is a rumor, the authors propose two greedy strategies based on the set of monitors that received the information. A first strategy tries to assign as many positive monitors as possible to each source, producing this way large greedy information propagation trees. To solve this problem, a second strategy estimates the disparity between actual propagation trees and the ones constructed by the above greedy strategy. To evaluate the proposed approach, a case study involving a real social network crawled from Twitter is reported. Using the experimental data, the authors evaluate how accurately LR can classify rumor and nonrumors. The proposed approach shows good potential to help users in identifying rumors and their sources.

Gupta et al.[97] propose an approach for the assessment of the credibility of news events on Twitter. Starting from the approach described by Castillo et al.,[79] they introduce some new features to improve it. The authors then claim that this classification-based approach is neither entity- nor network-aware, because events are described by features originally related to tweets and users. To address this issue, the authors propose an approach constituted by two modules: (1) a *Basic Credibility Analyzer*, namely

BasicCA, and (2) an *Event Graph Optimization Credibility Analyzer*, namely *EventOptCA*. *BasicCA* acts on a graph constituted by users, tweets, and events. At each iteration, each node shares its credibility value (learned from a classifier) with its neighbors only. Since the assumption of *BasicCA* is that credible entities are strictly connected, every iteration helps in mutually enhancing the credibility of genuine entities and reducing the credibility of nongenuine ones, via propagation. *EventOptCA* enhances *BasicCA* by supposing that similar events have similar credibility scores. Thus, it performs event credibility updates on a graph of events whose edges are weighted with event similarity values. The experiments are conducted using 457 news events extracted from two tweet feed datasets: a first dataset from Castillo et al.,[79] and another crawled datases, whose events are manually labeled as social gossip or news, using 250 of the news events (of which 167 labeled as credible) in their study. On an average, the proposed approach outperforms the classifier-based approach discussed by the authors.

The work just described is taken as a reference by Jin et al.,[83] who first discuss its limitations. In fact, the authors believe that it is not always true that credible users provide credible tweets with a high probability; furthermore, they are convinced that considering an event as a whole as only constituted by one kind of information (i.e., genuine or fake) is debatable. For this reason, they propose a hierarchical credibility propagation network with three layers: a message layer, a subevent layer and an event layer. The assumption behind this choice is that the hierarchical structure of message to subevent, and subevent to event, can model their relations and the process of credibility propagation; furthermore, with a subevent layer, deeper semantic information can be revealed for an event. Credibility propagation is modeled as a graph optimization problem. To validate the effectiveness of the proposed model, two datasets on Sina Weibo (the leading microblogging platform in China), were collected: one with random fake news in a year and truthful news at the same time; another with both fake and truthful news related to the same topic. Experiments on both datasets showed the effectiveness of the proposed model in terms of both accuracy and f-score improvements with respect to baseline methods.

Zhao et al.[98] consider the issue of automatically estimating the trustworthiness of messages and users in Twitter for a specific topic domain. In particular, the authors consider the relationships between users/tweets and multiple characteristics (i.e., textual, spatial, and temporal features) connected to them. First, the approach evaluates the trustworthiness of each tweet and each user. To do this, the authors evaluate similarity between features, under the assumption that a candidate tweet (and the user who wrote it) is considered trustworthy if its features do not conflict with the features of trustworthy news. Then, by means of four propagation rules defined on the social graph, the trustworthiness of tweets and users is refined and propagated. The evaluation of the similarity-based trust evaluation method is based on two datasets: a manually labeled set, and the dataset provided by Castillo et al.[79]; both datasets are also employed to identify emerging events. Based on the resulting precision and f-score values, this method outperforms classification-based supervised learning approaches.

Survey-Based Studies

These studies do not propose algorithms for the automatic or semi-automatic classification of fake news or spammers in Social Media. They are valuable since they employ survey-based studies, i.e., involving a representative sample of people who are asked to fill out questionnaires and/or to execute some tasks, to investigate the contribution that credibility factors have in the overall process of credibility judgment by users in Twitter and microblogs in general.

Morris et al.[99] present survey results regarding users perceptions of tweet credibility. The authors conduct some experiments with a limited number of persons involved, where several features are systematically manipulated to assess their impact on credibility ratings of tweets. The outcome of this research work is that the analysis of the content alone does not allow users to judge tweet credibility; instead, they are more influenced by heuristics such as user names or tweet topics when making credibility assessments, and less by profile picture images. This latter result has been recently denied by Kang et al.[100]; in their work, the authors identify the important cues that contribute to information being perceived as credible in microblogging, by reporting on a demographic survey followed by a user experiment with 101 people. This study investigates how such cues are portable across different microblogs platforms, i.e., Twitter and Reddit. The results of this study show that meta-data and image type elements are, in general, the strongest influencing factors in credibility assessment.

The fact that multiple kinds of features considered together leads to better results in evaluating perceived credibility has been experimentally confirmed by automatic classification-based approaches previously described, both for fake news and fake reviews

detection. An analysis of the distribution of the different types of salient features in Twitter is discussed by O'Donovan et al.,[101] together with the importance of considering the context in which these features are employed. A theoretical proposal for a contextual credibility approach has been illustrated by AlMansour et al.[82]; it aims at examining the effects that culture, topic variations, language, and so on, have on assessing credibility. Cultural differences and other contextual aspects affecting credibility perceptions have been studied by Yang et al.,[102] who point out the differences between China and United States users in rating the credibility of tweets. From this study, it emerges that Chinese people are apparently more willing to trust microblogs as information diffusion media, and to accept as reliable the content produced by anonymous or pseudo-anonymous sources.

Differently from the previous ones, the work by Sikdar et al.[103] illustrates that survey-based methods can be extremely noisy and that results may vary from survey to survey. In large-scale online user surveys, e.g., AMT, people providing credibility ratings usually ignore message senders and other external information such as the underlying network context, and they provide in this way biased credibility ratings. For this reason, the authors make a proposal to construct a stable ground truth, collecting two datasets on the same topic, but from different perspectives. To do this, two different surveys are provided to users, in which subjects are confronted with different information about the same tweets. This allows to test to which degree the credibility judgments across different surveys are comparable, and how the survey-based method influences the results. The authors also consider two different ways to quantify retweets, overall and at the time of the message, capturing the importance of the message at two different time granularities. This proposal represents an interesting contribution to the problem of generating a consensual and reliable ground truth that, as we have already discussed in the section illustrating opinion spam detection approaches, remains one of the most significant obstacles in evaluating the effectiveness of methods for credibility assessment in different contexts.

Discussion

As in the case of opinion spam detection in review sites, different approaches have been proposed to identify misinformation in microblogging. In this medium, the messages that are spread between users can correspond both to conversation items (e.g., chats, personal comments, and gossips) and news items (e.g., trending topics). Due to the influence that this second category of messages can have on a broader community, the research proposals have focused on approaches that allow the identification of fake news and spammers/group spammers diffusing them. The techniques that have been proposed so far do not differ substantially from those proposed for fake review detection; they focus on multiple kinds of characteristics connected to users, messages, topics, social graphs, and they apply some machine-learning techniques or propagation-based approaches to identify pieces or sources of misinformation. With respect to the opinion spam scenario, it is easier in fake news detection to evaluate the effectiveness of the proposed approaches, since the credibility of news and events can be verified (at least a posteriori) by experts and traditional media. Nevertheless, also in this case, in many approaches the volume of the datasets used for experimental evaluations remains rather low. Table 2 summarizes the research works that have been proposed so far by illustrating, for each approach, its focus and the category to which it belongs.

Credibility Assessment of Online Health Information

According to the Pew Research Center[b] (a nonpartisan 'fact tank' that informs the public about the issues, attitudes and trends shaping America and the world), about 59% of Americans have searched the Web frequently for health information in 2013. In particular, the 35% of U.S. adults declared to have used online health information to try to figure out their medical condition or that of someone they know. In addition to this, the 41% of the so called 'online diagnosers' reported their condition confirmed by a clinician.

Similarly, a recent EU-wide survey[104] showed that Europeans too are very sensitive to online information about health and healthy lifestyles. Six of 10 EU citizens go online when looking for health information, 90% of those said that the Web helped them to improve their knowledge about health-related topics.

Despite this interest, and the huge amount of health information available online, the extent to which people are capable to benefit from this information largely depends on their level of *health literacy*. According to the World Health Organization,[105] 'health literacy represents the cognitive and social skills which determine the motivation and the ability of individuals to gain access to, understand and use information in ways which promote and maintain

TABLE 2 | Summarization of the Proposed Approaches for Fake News Detection in Microblogging

Authors	Year	Focus	Approach
Mendoza et al.[95]	2010	Rumor propagation	Propagation-based
Castillo et al.[79]	2011	Trending topic credibility	Classification-based
Castillo et al.[84]	2012	Trending topic credibility	Classification-based
Gupta and Kumaraguru[88]	2012	Event credibility	Classification-based
Gupta et al.[97]	2012	Event credibility	Propagation-based
Kang et al.[87]	2012	Tweet and source credibility	Classification-based
Morris et al.[99]	2012	Perceived credibility factors	Survey-based
O'Donovan et al.[101]	2012	Perceived credibility and context	Survey-based
Seo et al.[96]	2012	Rumor and source credibility	Propagation-based
Galán-Garca et al.[93]	2013	Troll detection	Classification-based
Sikdar et al.[103]	2013	Ground truth construction	Survey-based
Yang et al.[102]	2013	Perceived credibility and context	Survey-based
AlMansour et al.[82]	2014	Perceived credibility and context	Survey-based
Gupta et al.[92]	2014	Tweet credibility	Classification-based
Jin et al.[83]	2014	Tweet and source credibility	Propagation-based
Gupta and Kaushal[94]	2015	Spammer detection	Classification-based
Kang et al.[100]	2015	Perceived credibility factors	Survey-based
Zhao et al.[98]	2016	Topic-focused tweet credibility	Propagation-based

good health.' More recently, based on 17 different definitions identified in the literature, the following definition has been provided by Sørensen et al.[106]: 'Health literacy is linked to literacy and entails people's knowledge, motivation and competencies to access, understand, appraise, and apply health information in order to make judgments and take decisions in everyday life concerning healthcare, disease prevention and health promotion to maintain or improve quality of life during the life course.' This definition is more focused on the *individual* with respect to the previous one, which is more centered on the public health perspective. In fact, over the years, a socially contextualized view of individuals has emerged, and the health promotion has shifted from simply encouraging people to adopt healthy behaviors and avoid unhealthy ones, to have more control over health, as single individuals and communities.[107]

Being able of assessing the reliability, validity, and in general the credibility of health information can be considered one of the aspects connected to health literacy.[108] As it may be easily guessed, the health information that individuals find online is likely used to assess and diagnose personal diseases; for this reason, inaccurate or misleading medical information would cause negative and harmful consequences for users' health.[16,109] In addition to this, if online health information is perceived as credible, people will invest more time in searching and using different types of online health-related services, avoiding to use a significant amount of cognitive efforts and knowledge that would represent a barrier for health information usage.[110]

In the last years, the environments in which people have searched for and/or consumed online information related to health has dramatically changed. From traditional Web search engines and specialized healthcare platforms, users has moved to Social Media platforms,[111,112] appreciating the fact that in this kind of media it is possible to exchange personal experience about their own medical or health-related issues. In this changing context, it is necessary to briefly recall that credibility has traditionally been defined as the believability of information messages and sources, as *perceived* by the information receiver.[113] Thus, depending on the environment, the way in which users can perceive and evaluate both the health source and message credibility changes. Before the advent of technologies related to the Web 2.0, the focus of research studies in this field was mainly on the investigation of the credibility of Websites spreading medical and health-related information. Nowadays, with the rapid diffusion of Social Media and the increasing volume of health information shared in these applications, new open

issues and a new (and very little explored) research field have emerged.

Health Information Credibility in Websites

The study of the reliability of healthcare Websites and the credibility of the information they contain has been investigated especially between 2000 and 2010, and it has concerned both patients and the health sector.[114,115] In particular, the content of health-related Websites is potentially unsafe if its quality has not been properly addressed.[116]

Over the years, researchers have demonstrated that there are several Website features that can boost the perceived credibility of information in the healthcare context, such as a professional-appearing interface design, the ease of navigation, the presence of endorsements, etc.[113] In their systematic review of health Website credibility evaluation,[117] Eysenbach et al. show that, among the characteristics that influence credibility, there are: *accuracy*, i.e., the degree of concordance of the information provided with generally accepted medical practice, *completeness*, i.e., the portion of topics covered by the Website, *readability*, and *design*, i.e., the visual appearance. The article by Freeman and Spyridakis[109] is one of the first studies summarizing and analyzing in a complete way the factors influencing the credibility of information in health-related Websites, and the review works by Metzger,[3] and Rieh and Danielson[118] provide a more general view on Website credibility, discussing at the same time some health-related issues.

In a more recent study, Rains and Donnerstein Karmikel[119] examine the relationship between Websites' structural features and message characteristics, in terms of perceived credibility. In addition, in assessing Website credibility, they study the impact of individuals' navigation orientation, i.e., searching or surfing. As already proved by previous state-of-the-art approaches, the authors show that structural features (e.g., a street address, privacy policy statements, third-party endorsements, and the inclusion of a navigation menu or images) are effectively related to perceptions of Website credibility. Furthermore, they illustrate how characteristics connected to the message (e.g., statistics, quotes, and identification of authorship) are positively associated with the evaluation of a given health topic in terms of credibility. An interesting outcome of this work is that these characteristics may have an influence beyond the particular Website on which they are presented, and may be extended to individuals' credibility perceptions about a health topic in general. This represents an interesting link to the study of health-related information credibility in Social Media, where the characteristics of the medium for information sharing have a lower impact in the perceived credibility assessment, especially with respect to the information source.[120]

Health Information Credibility in Social Media

In general, individuals are inclined to participate in Social Media to share their stories, experience, and knowledge, since social interactions offer them a sense of satisfaction to be a member of a community.[121] This is even more true in such a special and personal context like the one represented by online communities related to health. These communities can either be small or large groups of individuals who share the same health problems, as well as professional healthcare services allowing people to interact online.[26] This represents for users the possibility to have an immediate satisfaction of their needs, since they can share their knowledge and symptoms; users can post/read reviews about health products, medicines, and doctors, thus forming online support groups or self-help groups. Social Media also constitutes a valuable opportunity for the health sector to improve services.[25] In fact, using Social Media, health care providers can post health information not only as text, but also in various accessible forms, such as images and videos. This way, Social Media empowers health care consumers,[122] by providing them with immediate access to an incredible amount of health information and a variety of perspectives on health topics,[112] which they would never have been able to access before, since owned by health care providers.

It is immediately clear to the reader how these benefits also bring with them a particularly important issue. The direct access to medical information and the disappearance of experienced intermediaries, makes the problem of verifying the credibility of the obtained information a key factor in the healthcare scenario. Here, more than in other environments, a careful evaluator is necessary to affirm that the posted information is trustworthy, and not dangerous for the patient.[16] Despite the importance of this problem, and the fact that it has been previously addressed in many research fields such as psychology and social sciences,[115,123–125] few are the approaches that have been so far proposed in computer science to assess, automatically or semi-automatically, the credibility of health-related information in Social Media. With respect to healthcare Websites, where erroneous comments can be corrected by third parties,[126] and some structural features can be positively connected to the credibility of the Website

itself,[119] online communities are characterized by different features and strategies connected to the evaluation of information credibility. It is possible, for example, to provide ratings with respect to pieces and sources of information, as well as to analyze features connected to both user profiles and other related information,[26] and to the UGC.

The paper by Oh et al.[127] investigates the issue of quality of online health answers in social question answering (QA) systems, considering also source credibility. This paper presents a preliminary user study involving 40 participants who have rated 400 health answers each with a 5-point Likert scale according to 10 evaluation criteria. The paper highlights the positive role that experts play in helping users and patients to find reliable health information in Social Media. Another preliminary work by Syn and Kim[128] addresses the impact of information sources on users' credibility perceptions. The paper discusses in particular the way in which credibility affects, in Facebook, the activities related to information generation and usage by users. According to the authors, these activities include the actions that people do to fill the gaps between needs and solutions in problematic situations connected to health, e.g., search and retrieval, browsing, monitoring, and so on. The approach is based on an online survey study, involving a limited number of people. The findings of this work include that, although users perceive certain sources of health information as unreliable, they still access them to find information; this finding also emerged in prior work.[129] The paper also shows that the fact of identifying on Facebook the sources of information more easily with respect to traditional Websites, has a positive impact on the users' perception of information credibility. Lederman et al.[130] provide qualitative and quantitative studies on how users assess the credibility of UGC in Online Health Forums (OHF), with particular reference to source and message characteristics. 159 participants were recruited using the AMT, and the analysis was performed by requesting these participants to complete a 10-min online survey on their credibility perceptions. From this work, it emerges that two kinds of information are generated in OHF: scientific information and experimental information. When a message explicitly refers to external scientific content, the health literacy of users becomes less important, and they are more likely to depend on objective standards; instead, when dealing with subjective or conflicting experimental information, users are more willing to focus on crowd consensus.[131] Ma and Atkin[120] in their recent work provide the first complete meta-analysis of perceived credibility of online user-generated health information. The aim is to synthesize the findings of past research about the source and content credibility relationship. As also illustrated in the above studies,[128,129] what emerges is that the original sources of online health information generally do not have a great impact on the perceived information credibility, especially in Social Media. In fact, when the health-related content is generated by laypersons, it maintains the potential of being perceived as highly credible. This does not happen in traditional Websites, where only the health information created by experts is perceived as highly credible. This means that in the social-oriented health context, individuals trust the content generated by laypersons, due to a particular characteristic of this kind of environments, i.e., *homophily*.[132] When the source has low credibility, homophily influences the credibility perceptions that users have on information. As seen so far, several are the works in the literature that have made an analysis of the factors that influence the perception of health-related information credibility in Social Media. However, few are the approaches that have been proposed up to now to assess the credibility of health information in an automatic or semi-automatic way. Abbasi et al.[133] propose a focused crawling method, *ssmCredible*, to collect credible medical content from Websites, forums, blogs, and social networking sites, by considering source and sentiment features pertaining the medical content. For a given candidate URL, the system computes a credibility score that is used both to rank candidate credible URLs, and to filter out non-credible URLs; the technique is based on a graph propagation algorithm that leverages credibility information from various online medical databases. Each credible URL is then evaluated in terms of relevance: text classifiers are applied to its parent pages, by considering linguistic features (i.e., words and *n*-grams incorporating medical and sentiment lexicons) learned from a set of manually labeled Web pages known to be relevant or irrelevant. Comparative evaluations have been performed against previous focused crawlers and baseline Web spiders,[134] showing the effectiveness of the proposed approach, in particular on blogs. Weitzel et al.[135] propose a method, *Reputation Rank*, to measure the source reputation in the health-related context by exploiting social interactions in Twitter. This approach focuses on retweeting as a spreading information mechanism, which is considered as a form of endorsement for both the message and the originating user. The authors model a *Retweet-Network* as a directed weighted graph, where users are the nodes and retweets are

the edges. A direct edge is generated if a user u_j retweets the content of another user u_i. Weights on edges represents the strength of trust ties, and are computed as the ratio between the number of retweets received by a user u_i from a user u_j, and the total number of u_i retweets in the network, plus a 'discount rate' representing the presence or absence of relationships (e.g., the 'follow' relationship) among users. The *Rank Reputation* approach proposed by the authors ranks the reputation of the sources of information based on a set of centrality measures[136] that exploit the *Retweet-Network* modeled as previously described.

The most complete work appeared so far in this domain is the one by Mukherjee et al.,[137] who propose a method for the automatic assessment of the credibility of both user-generated medical statements in users' posts and their authors. This approach considers both linguistic- and user-based features, since it focuses on the assumption that statement credibility strongly depends on linguistic objectivity, i.e., the general quality of the language in the user's post, and on user trustworthiness, represented by her/his status and engagement in the community (e.g., the number of questions and answers posted, the number of provided 'thanks,' etc.). These factors are modeled through an MRF model, where the random variables are constituted by users, their posts, and the medical statements contained within posts. Evaluations are made on 15,000 users and 2.8 million posts crawled from healthboards.com, and on a ground truth provided by a medical portal. The results produced by the application of the proposed model are compared with state-of-the-art baselines, illustrating an increase in accuracy with respect to SVM.

Discussion

With respect to opinion spam and fake news detection, a limited number of approaches have been proposed in the health-related context to assess the credibility of online information, especially in Social Media. Table 3 summarizes the main approaches that have been proposed so far; some of them consider the issue of Website credibility, while in the Social Media scenario the majority of them concerns the analysis of the characteristics influencing users' perception of credibility with respect to health information. Only few works propose concrete techniques to assess sources and/or information credibility in Social Media. This shows how this problem, although it is of paramount importance, has to be fully investigated yet.

CONCLUSIONS

In the post-truth era, Social Media often become the vehicle of the spread of misinformation and hoaxes. Individuals who must form their own opinion, and who are often locked inside their 'eco-chambers,' do not have the necessary instruments and cognitive abilities to assess the level of credibility of pieces and sources of information with which they come into contact. While this may not have serious consequences in the case of the spread of rumors of little importance, it may lead to serious problems when consumers are directed to choose products and services based on fake reviews, or when false news (e.g., political news) that influence public opinion are diffused, or when individuals run into inaccurate medical information which is likely to jeopardize their health. For this reason, the interest in studying possible ways of helping users in assessing the level of credibility of online information is particularly important, especially in the Social Web scenario. The

TABLE 3 | Summarization of the Proposed Approaches for Assessing Credibility of Online Health Information

Authors	Year	Focus	Contribution
Eysenbach et al.[117]	2002	Website credibility	Credibility perception analysis
Freeman and Spyridakis[109]	2004	Website credibility	Credibility perception analysis
Rains and Donnerstein Karmikel[119]	2009	Website credibility	Credibility perception analysis
Oh et al.[127]	2012	Credibility in social media	Credibility perception analysis
Abbasi et al.[133]	2013	Credibility in social media	Crawling credible UGC
Syn and Kim[128]	2013	Credibility in social media	Credibility perception analysis
Lederman et al.[130]	2014	Credibility in social media	Credibility perception analysis
Mukherjee et al.[137]	2014	Credibility in social media	UGC credibility assessment
Weitzel et al.[135]	2014	Credibility in social media	Source reputation assessment
Ma and Atkin[120]	2016	Credibility in social media	Credibility perception analysis

concept of credibility has been studied since Aristotle (fourth century BC), and more recently in many different research fields, such as communication, psychology, and social sciences. An interesting aspect connected to credibility is that, according to several studies, it represents a perceived quality of individuals, who form their opinion in a way to minimize their cognitive effort, especially online. In everyday life, people usually reduce the uncertainty about credibility based on information easy to obtain in the 'offline' world: the reputation of the source of information, the presence of trusted intermediaries such as experts and/or opinion leaders, personal trust based on first-hand experiences. Conversely, the digital realm is characterized by a process of 'disintermediation'; in this context, the multiplicity of sources involved in information dissemination, their possible anonymity, the absence of standards for information quality, the ease in manipulating and altering contents, the lack of clarity of the context, and the presence of many potential targets of credibility evaluation (i.e., the content, the source, and the medium), make credibility assessment a more complex task with respect to 'offline' media.

Due to the above reasons, in the last years numerous approaches in the literature have tackled the issue of the automatic assessment of information credibility, in particular in Social Media. Individuals' credibility perceptions can be evaluated in terms of multiple characteristics, which may be associated with sources of information, shared contents, and media across which information diffuses. In this survey, we have presented the more significant state-of-the-art proposals, which include: data-driven approaches, model-driven approaches, and approaches that exploit the graph-based structure connecting entities in social platforms to propagate credibility and trust. In particular, we have taken into consideration that the Social Web is characterized by multiple social applications, each of which is intended for different audiences and aims. This means that the solutions that have been proposed to assess credibility in Social Media change depending on the considered scenario. In this article, we have therefore considered some context where the spread of false information may generate serious consequences for the individuals involved. In particular, we have presented state-of-the-art approaches addressing three main tasks: (1) the detection of opinion spam in review sites, (2) the detection of fake news and spam in microblogging, and (3) the credibility assessment of online health information. Table 4 summarizes and compares the three above tasks with respect to their focus, proposed approaches, and open issues in the assessment of information credibility.

TABLE 4 | Summarization of the Focus, Proposed Approaches, and Open Issues of Each of the Three Scenarios Considered in the Credibility Assessment Issue

Assessment of the Credibility of Information in Social Media		
Opinions	News	Health Information
	Focus	
Fake review detection	Perceived credibility factors	Perceived credibility factors
Spammer detection	Rumor propagation	Health Website credibility
Group spammer detection	Spammer detection	Health-related UGC credibility assessment
	Tweet/Event/Trending topic credibility	
	Approaches	
Classification-based	Classification-based	Classification-based
Content-based	Multifeature-based	Multifeature-based
Behavioral-based	Propagation-based	Propagation-based
Multifeature-based	Graph-based	Graph-based (Crawling)
Propagation-based	Survey-based	Survey-based
Behavioral-based		
Graph-based		
	Open issues	
Almost impossible to hire human experts for assessing the credibility of 'opinions', such as reviews	The impact of the content alone is not sufficiently investigated	Only few automatic or semi-automatic approaches are proposed
Absence of convincing gold standard datasets	The definition of news, topic, event is not always rigorous	Higly context-dependent scenario
Small datasets are considered	Small datasets are considered	Possibly harmful scenario
		Small datasets are considered

As discussed in the article, and as it emerges from Table 4, among the important open issues characterizing the credibility assessment problem, the most urgent are the absence of predefined benchmarks and gold standard datasets (in particular, for fake review detection), and the problem of collecting and mining large amounts of data. In the big data era, this latter problem has not yet received the attention it deserves. Furthermore, only few approaches have been proposed to automatically assess the credibility of online health information, even if we think that this constitutes one of the most challenging areas of research in the coming years.

NOTES

[a] https://twitter.com
[b] http://www.pewresearch.org/

REFERENCES

1. Quattrociocchi W, Scala A, Sunstein CR. *Echo Chambers on Facebook*, June 13, 2016. Available: http://ssrn.com/abstract=2795110.
2. Self CS. Credibility. In: Salwen MB, Stacks DW, eds. *An Integrated Approach to Communication Theory and Research*. 2nd ed. Routledge: Taylor and Francis Group; 2008, 435–456. https://doi.org/10.4324/9780203887011.
3. Metzger MJ. Making sense of credibility on the web: models for evaluating online information and recommendations for future research. *J Am Soc Inf Sci Technol* 2007, 58:2078–2091.
4. Hovland CI, Lumsdaine AA, Fred D. *Experiments on Mass Communication*. Studies in Social Psychology in World War II, vol. 3. Princeton, NJ: Princeton University Press; 1949.
5. Hovland CI. Changes in attitude through communication. *J Abnorm Soc Psychol* 1951, 46:424.
6. Hovland CI, Janis IL, Kelley HH. *Communication and Persuasion*. New Haven, CT: Yale University Press; 1953.
7. Fogg BJ, Tseng H. The elements of computer credibility. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Pittsburgh, PA, USA, 15–20 May, 1999. New York, NY: ACM; 1999, 80–87.
8. Metzger MJ, Flanagin AJ, Eyal K, Lemus DR, McCann RM. Credibility for the 21st century: integrating perspectives on source, message, and media credibility in the contemporary media environment. *Ann Int Commun Assoc* 2003, 27:293–335.
9. Pornpitakpan C. The persuasiveness of source credibility: a critical review of five decades' evidence. *J Appl Soc Psychol* 2004, 34:243–281. https://doi.org/10.1111/j.1559-1816.2004.tb02547.x.
10. Liu Z. Perceptions of credibility of scholarly information on the web. *Inf Process Manage* 2004, 40:1027–1038.
11. Bernier CL. Second-hand knowledge. An inquiry into cognitive authority. Patrick Wilson. Greenwood Press, 1983. *J Am Soc Inf Sci* 1984, 35:254–255. https://doi.org/10.1002/asi.4630350410.
12. Hilligoss B, Rieh SY. Developing a unifying framework of credibility assessment: construct, heuristics, and interaction in context. *Inf Process Manage* 2008, 44:1467–1484. https://doi.org/10.1016/j.ipm.2007.10.001.
13. Wathen CN, Burkell J. Believe it or not: factors influencing credibility on the web. *J Am Soc Inf Sci Technol* 2002, 53:134–144.
14. Metzger MJ, Flanagin AJ. Credibility and trust of information in online environments: the use of cognitive heuristics. *J Pragmat* 2013, 59(Part B):210–220. https://doi.org/10.1016/j.pragma.2013.07.012.
15. Flanagin AJ, Metzger MJ. *Digital Media and Youth: Unparalleled Opportunity and Unprecedented Responsibility*. Cambridge, MA: MIT Press; 2008, 5–28. Available at: http://www.mitpressjournals.org/doi/pdfplus/10.1162/dmal.9780262562324.005. Accessed January 1, 2017.
16. Eysenbach G. Credibility of Health Information and Digital Media: New Perspectives and Implications for Youth. In: Metzger MM, Flanagin AJ, eds. *Digital Media, Youth, and Credibility*. Cambridge, MA: The MIT Press; 2008, 123–154.
17. Sundar SS. The main model: a heuristic approach to understanding technology effects on credibility. In: Metzger MJ, Flanagin AJ, eds. *Digital Media, Youth, and Credibility*. Cambridge, MA: The MIT Press; 2008, 73–100.
18. Eysenbach G, Köhler C. How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. *BMJ* 2002, 324:573–577.
19. Kaplan AM, Haenlein M. Users of the world, unite! The challenges and opportunities of social media. *Bus Horiz* 2010, 53:59–68. DOI: https://doi.org/10.1016/j.bushor.2009.09.003.

20. Warneken F, Tomasello M. Altruistic helping in human infants and young chimpanzees. *Science* 2006, 311:1301–1303.
21. Olson KR, Spelke ES. Foundations of cooperation in young children. *Cognition* 2008, 108:222–231.
22. Warneken F, Tomasello M. The roots of human altruism. *Br J Psychol* 2009, 100:455–471.
23. Moens M-F, Li J, Chua T-S, eds. *Mining User Generated Content, Social Media and Social Computing*. Boca Raton, FL: Chapman and Hall/CRC; 2014.
24. Safko L. *The Social Media Bible: Tactics, Tools, and Strategies for Business Success*. 2nd ed. Hoboken, NJ John Wiley & Sons Publishing; 2010. ISBN: 0470623977.
25. Hajli MN. Developing online health communities through digital media. *Int J Inf Manage* 2014, 34:311–314.
26. Hajli MN, Sims J, Featherman M, Love PE. Credibility of information in online communities. *J Strateg Market* 2015, 23:238–253.
27. Lang A. The limited capacity model of mediated message processing. *J Commun* 2000, 50:46–70.
28. Liu B. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer-Verlag Berlin Heidelberg Springer Science & Business Media; 2007.
29. Jindal N, Liu B. Review spam detection. In: *Proceedings of the 16th International Conference on World Wide Web*, Banff, Alberta, Canada, 8–12 May, 2007. New York, NY: ACM; 2007, 1189–1190.
30. Jindal N, Liu B. Opinion spam and analysis. In: *Proceedings of the 2008 International Conference on Web Search and Data Mining*, Palo Alto, CA, USA, 11–12 February, 2008. New York, NY: ACM; 2008, 219–230.
31. Heydari A, Tavakoli M, Salim N, Heydari Z. Detection of review spam: a survey. *Expert Syst Appl* 2015, 42:3634–3642.
32. Crawford M, Khoshgoftaar TM, Prusa JD, Richter AN, Al Najada H. Survey of review spam detection using machine learning techniques. *J Big Data* 2015, 2:23. https://doi.org/10.1186/s40537-015-0029-9.
33. Li F, Huang M, Yang Y, Zhu X. Learning to identify review spam. In: *IJCAI Proceedings—International Joint Conference on Artificial Intelligence*, vol 22, Barcelona, Catalonia, Spain, 16–22 July, 2011. Menlo Park, CA: AAAI Press; 2011, 2488–2493.
34. Ott M, Choi Y, Cardie C, Hancock JT. Finding deceptive opinion spam by any stretch of the imagination. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, vol 1, Portland, OR, USA, 19–24 June, 2011. Stroudsburg, PA: Association for Computational Linguistics; 2011, 309–319.
35. Ott M, Cardie C, Hancock J. Estimating the prevalence of deception in online review communities. In: *Proceedings of the 21st International Conference on World Wide Web*, Lyon, France, 16–20 April, 2012. New York, NY: ACM; 2012, 201–210.
36. Banerjee S, Chua AY. Applauses in hotel reviews: genuine or deceptive?. In: *Science and Information Conference* (*SAI*), London, UK, 27–29 August, 2014. IEEE; 2014, 938–942.
37. Fusilier DH, Montes-y Gómez M, Rosso P, Cabrera RG. Detection of opinion spam with character n-grams. In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer; 2015, 285–294.
38. Hengeveld K. Parts of speech. In: Fortescue M, Harder P, Kristoffersen L, eds. *Layered Structure and Reference in a Functional Perspective*. Amsterdam: John Benjamins; 1992, 29–55.
39. Pennebaker JW, Francis ME, Booth RJ. *Linguistic Inquiry and Word Count: Liwc 2001*, vol. 71. Mahway, NJ: Lawrence Erlbaum Associates; 2001, 2001.
40. Mukherjee A, Venkataraman V, Liu B, Glance N. Fake review detection: classification and analysis of real and pseudo reviews. Technical Report No. UIC-CS-03-2013, 2013.
41. Mukherjee A, Venkataraman V, Liu B, Glance NS. What yelp fake review filter might be doing?. In: *ICWSM*, 2013.
42. Sun H, Morales A, Yan X. Synthetic review spamming and defense. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY: ACM; 2013, 1088–1096.
43. Croft B, Lafferty J. *Language Modeling for Information Retrieval*, vol. 13. Dordrecht: Springer Science & Business Media; 2013.
44. Lai CL, Xu KQ, Lau RYK, Li Y, Jing L. Toward a language modeling approach for consumer review spam detection. In: *2010 I.E. 7th International Conference on E-Business Engineering*, 2010, 1–8.
45. Kullback S, Leibler RA. On information and sufficiency. *Ann Math Stat* 1951, 22:79–86.
46. Lau RYK, Liao SY, Kwok RC-W, Xu K, Xia Y, Li Y. Text mining and probabilistic language modeling for online review spam detection. *ACM Trans Manage Inf Syst* 2012, 2:25:1–25:30. https://doi.org/10.1145/2070710.2070716.
47. Mukherjee A, Liu B, Wang J, Glance N, Jindal N. Detecting group review spam. In: *Proceedings of the 20th International Conference Companion on World Wide Web*. New York, NY: ACM; 2011, 93–94.
48. Li X, Liu B. Learning to classify texts using positive and unlabeled data. In: *IJCAI*, vol 3. 2003, 587–592.

49. Ren Y, Ji D, Zhang H. Positive unlabeled learning for deceptive reviews detection. In: *EMNLP*, 2014, 488–498.
50. Yafeng R, Lan Y, Donghong J. Deceptive reviews detection based on language structure and sentiment polarity. *J Front Comput Sci Technol* 2014, 8:313–320.
51. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. *J Mach Learn Res* 2003, 3:993–1022.
52. Fusilier DH, Móntes-y-Gómez M, Rosso P, Cabrera RG. Detecting positive and negative deceptive opinions using pu-learning. *Inf Process Manage* 2015, 51:433–443. DOI: https://doi.org/10.1016/j.ipm.2014.11.001.
53. Liu B, Lee WS, Yu PS, Li X. Partially supervised classification of text documents. In: *ICML* (Citeseer), vol 2, 2002, 387–394.
54. Abbasi A, Zhang Z, Zimbra D, Chen H, Nunamaker JF Jr. Detecting fake websites: the contribution of statistical learning theory. *MIS Q* 2010, 34:435–461.
55. Li F, Huang M, Yang Y, Zhu X. Learning to identify review spam. In: *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, IJCAI'11, vol 3, Barcelona, Catalonia, Spain. AAAI Press; 2011, 2488–2493, ISBN 978-1-57735-515-1.
56. Li H, Liu B, Mukherjee A, Shao J. Spotting fake reviews using positive-unlabeled learning. *Comput Sist* 2014, 18:467–475.
57. Luca M, Zervas G. Fake it till you make it: reputation, competition, and yelp review fraud. *Manage Sci* 2016, 62:3412–3427.
58. Viviani M, Pasi G. Quantifier guided aggregation for the veracity assessment of online reviews. *Int J Intell Syst* 2016, 32:481–501.
59. Lim E-P, Nguyen V-A, Jindal N, Liu B, Lauw HW. Detecting product review spammers using rating behaviors. In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. New York, NY: ACM; 2010, 939–948.
60. Fei G, Mukherjee A, Liu B, Hsu M, Castellanos M, Ghosh R. Exploiting burstiness in reviews for review spammer detection. In: *ICWSM*, vol 13. 2013, 175–184.
61. Botev ZI, Grotowski JF, Kroese DP, et al. Kernel density estimation via diffusion. *Ann Stat* 2010, 38:2916–2957.
62. Cross GR, Jain AK. Markov random field texture models. *IEEE Trans Pattern Anal Mach Intell* 1983, PAMI-5:25–39.
63. Metzler D, Croft WB. A Markov random field model for term dependencies. In: *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY: ACM; 2005, 472–479.
64. Murphy KP, Weiss Y, Jordan MI. Loopy belief propagation for approximate inference: an empirical study. In: *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc.; 1999, 467–475.
65. Mukherjee A, Liu B, Glance N. Spotting fake reviewer groups in consumer reviews. In: *Proceedings of the 21st International Conference on World Wide Web*. New York, NY: ACM; 2012, 191–200.
66. G. Wang, S. Xie, B. Liu, and S. Y. Philip. Review graph based online store review spammer detection. In: *2011 I.E. 11th International Conference on Data Mining*. IEEE; 2011, 1242–1247.
67. Wang G, Xie S, Liu B, Yu PS. Identify online store review spammers via social review graph. *ACM Trans Intell Syst Technol* 2012, 3:61.
68. Akoglu L, Chandy R, Faloutsos C. Opinion fraud detection in online reviews by network effects. In: *ICWSM*, vol 13; 2013, 2–11.
69. Ye J, Akoglu L. Discovering opinion spammer groups by network footprints. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer; 2015, 267–282.
70. Choo E, Yu T, Chi M. *Detecting Opinion Spammer Groups Through Community Discovery and Sentiment Analysis*. Cham: Springer International Publishing; 2015, 170–187. https://doi.org/10.1007/978-3-319-20810-7_11. ISBN: 978-3-319-20810-7.
71. Wang Z, Hou T, Song D, Li Z, Kong T. Detecting review spammer groups via bipartite graph projection. *Comput J* 2015, 59:bxv068.
72. Borgelt C. Frequent item set mining. *WIREs Data Mining Knowl Discov* 2012, 2:437–456.
73. Li H, Chen Z, Mukherjee A, Liu B, Shao J. Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns. In: *International AAAI Conference on Web and Social Media*, 2015, 634–637.
74. Kaplan AM, Haenlein M. The early bird catches the news: nine things you should know about micro-blogging. *Bus Horiz* 2011, 54:105–113.
75. Kaplan AM, Haenlein M. Users of the world, unite! The challenges and opportunities of social media. *Bus Horiz* 2010, 53:59–68.
76. Harcup T, O'Neill D. What is news? Galtung and ruge revisited. *J Stud* 2001, 2:261–280.
77. Java A, Song X, Finin T, Tseng B. Why we twitter: understanding microblogging usage and communities. In: *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*. New York, NY: ACM; 2007, 56–65.

78. Naaman M, Boase J, Lai C-H. Is it really about me?: message content in social awareness streams. In: *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*. New York, NY: ACM; 2010, 189–192.

79. Castillo C, Mendoza M, Poblete B. Information credibility on twitter. In: *Proceedings of the 20th International Conference on World Wide Web*. New York, NY: ACM; 2011, 675–684.

80. Kwak H, Lee C, Park H, Moon S. What is Twitter, a social network or a news media?. In: *Proceedings of the 19th International Conference on World Wide Web*. New York, NY: ACM; 2010, 591–600.

81. Rubin VL, Chen Y, Conroy NJ. Deception detection for news: three types of fakes. *Proc Assoc Inf Sci Technol* 2015, 52:1–4.

82. AlMansour AA, Brankovic L, Iliopoulos CS. A model for recalibrating credibility in different contexts and languages—a twitter case study. *Int J Digital Inf Wirel Commun* 2014, 4:53–62.

83. Jin Z, Cao J, Jiang Y-G, Zhang Y. News credibility evaluation on microblog with a hierarchical propagation model. In: *2014 I.E. International Conference on Data Mining*. IEEE; 2014, 230–239.

84. Castillo C, Mendoza M, Poblete B. Predicting information credibility in time-sensitive social media. *Internet Res* 2012, 23:560–588. https://doi.org/10.1108/IntR-05-2012-0095.

85. Mathioudakis M, Koudas N. TwitterMonitor: trend detection over the Twitter stream. In: *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*. New York, NY: ACM; 2010, 1155–1158.

86. Bhargava N, Sharma G, Bhargava R, Mathuria M. Decision tree analysis on j48 algorithm for data mining. *Proc Int J Adv Res Comput Sci Softw Eng* 2013, 3:1114–1119.

87. Kang B, O'Donovan J, Höllerer T. Modeling topic specific credibility on twitter. In: *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*. New York, NY: ACM; 2012, 179–188.

88. Gupta A, Kumaraguru P. Credibility ranking of tweets during high impact events. In: *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*. New York, NY: ACM; 2012, 2.

89. Cao Y, Xu J, Liu T-Y, Li H, Huang Y, Hon H-W. Adapting ranking SVM to document retrieval. In: *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY: ACM; 2006, 186–193.

90. Trotman A. Learning to rank. *Inf Retr* 2005, 8:359–381.

91. Järvelin K, Kekäläinen J. Cumulated gain-based evaluation of ir techniques. *ACM Trans Inf Syst* 2002, 20:422–446.

92. Gupta A, Kumaraguru P, Castillo C, Meier P. Tweetcred: A real-time Web-based system for assessing credibility of content on Twitter. In: *Proceedings of 6th International Conference on Social Informatics* (*SocInfo*), Barcelona, Spain, 2014.

93. Galán-Garca P, de la Puerta JG, Gómez CL, Santos I, Bringas PG. Supervised machine learning for the detection of troll profiles in twitter social network: application to a real case of cyberbullying. In: *International Joint Conference SOCO'13-CISIS'13-ICEUTE'13*. Springer; 2014, 419–428.

94. Gupta A, Kaushal R. Improving spam detection in online social networks. In: *2015 International Conference on Cognitive Computing and Information Processing* (*CCIP*). IEEE; 2015, 1–6.

95. Mendoza M, Poblete B, Castillo C. Twitter under crisis: can we trust what we RT? In: *Proceedings of the First Workshop on Social Media Analytics*. New York, NY: ACM; 2010, 71–79.

96. Seo E, Mohapatra P, Abdelzaher T. Identifying rumors and their sources in social networks. In: *SPIE Defense, Security, And Sensing*. International Society for Optics and Photonics; 2012, 83891I–83891I.

97. Gupta M, Zhao P, Han J. Evaluating event credibility on twitter. In: *SDM*. SIAM; 2012, 153–164.

98. Zhao L, Hua T, Lu C-T, Chen I-R. A topic-focused trust model for twitter. *Comput Commun* 2016, 76:1–11. DOI: https://doi.org/10.1016/j.comcom.2015.08.001.

99. Morris MR, Counts S, Roseway A, Hoff A, Schwarz J. Tweeting is believing?: understanding microblog credibility perceptions. In: *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, Seattle, Washington, USA. New York, NY: ACM; 2012, 441–450, ISBN 978-1-4503-1086-4. DOI: https://doi.org/10.1145/2145204.2145274.

100. Kang B, Höllerer T, O'Donovan J. Believe it or not? analyzing information credibility in microblogs. In: *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2015. New York, NY: ACM; 2015, 611–616.

101. O'Donovan J, Kang B, Meyer G, Höllerer T, Adalii S. Credibility in context: an analysis of feature distributions in twitter. In: *2012 International Conference on Privacy, Security, Risk and Trust* (*PASSAT*), *2012 International Conference on Social Computing (SocialCom)*. IEEE; 2012, 293–301.

102. Yang J, Counts S, Morris MR, Hoff A. Microblog credibility perceptions: comparing the USA and China. In: *Proceedings of the 2013 Conference on*

Computer Supported Cooperative Work. New York, NY: ACM; 2013, 575–586.

103. Sikdar S, Kang B, O'Donovan J, Höllerer T, Adah S. Understanding information credibility on twitter. In: *2013 International Conference on Social Computing* (*SocialCom*). IEEE; 2013, 19–24.
104. European Commission. Flash Eurobarometer 404—TNS Political & Social. Technical Report. European Commission; 2014.
105. World Health Organization, Kickbusch I, Nutbeam D. *Health Promotion Glossary*. Geneva: World Health Organization; 1998.
106. Sørensen K, Van den Broucke S, Fullam J, Doyle G, Pelikan J, Slonska Z, Brand H. Health literacy and public health: a systematic review and integration of definitions and models. *BMC Public Health* 2012, 12:1.
107. Stokols D. Translating social ecological theory into guidelines for community health promotion. *Am J Health Promot* 1996, 10:282–298.
108. Chinn D. Critical health literacy: a review and critical analysis. *Soc Sci Med* 2011, 73:60–67.
109. Freeman KS, Spyridakis JH. An examination of factors that affect the credibility of online health information. *Tech Commun* 2004, 51:239–263.
110. Xiao N, Sharman R, Rao HR, Upadhyaya S. Factors influencing online health information search: an empirical analysis of a national cancer-related survey. *Decis Support Syst* 2014, 57:417–427.
111. De Choudhury M, Morris MR, White RW. Seeking and sharing health information online: comparing search engines and social media. In: *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*. New York, NY: ACM; 2014, 1365–1376.
112. Li Y, Wang X, Lin X, Hajli M. Seeking and sharing health information on social media: a net valence model and cross-cultural comparison. *Technol Forecast Social Change*. In press.
113. Metzger MJ, Flanagin AJ. Online Health Information Credibility. In: Thompson TL, ed. *Encyclopedia of Health Communication*. Thousand Oaks, CA: SAGE; 2011, 976–978.
114. Rice RE. Influences, usage, and outcomes of internet health information searching: multivariate results from the pew surveys. *Int J Med Inform* 2006, 75:8–28.
115. Adams SA. Revisiting the online health information reliability debate in the wake of web 2.0: an interdisciplinary literature and website review. *Int J Med Inform* 2010, 79:391–400.
116. O'Grady L, Wathen CN, Charnaw-Burger J, Betel L, Shachak A, Luke R, Hockema S, Jadad AR. The use of tags and tag clouds to discern credible content in online health message forums. *Int J Med Inform* 2012, 81:36–44.
117. Eysenbach G, Powell J, Kuss O, Sa E-R. Empirical studies assessing the quality of health information for consumers on the world wide web: a systematic review. *JAMA* 2002, 287:2691–2700.
118. Rieh SY, Danielson DR. Credibility: a multidisciplinary framework. *Annu Rev Inf Sci Technol* 2007, 41:307–364.
119. Rains SA, Karmikel CD. Health information-seeking and perceptions of website credibility: examining web-use orientation, message characteristics, and structural features of websites. *Comput Hum Behav* 2009, 25:544–553, including the Special Issue: State of the Art Research into Cognitive Load Theory. DOI: https://doi.org/10.1016/j.chb.2008.11.005.
120. Ma TJ, Atkin D. User generated content and credibility evaluation of online health information: a meta analytic study. *Telemat Inform* 2016, 34:472–486.
121. Casaló LV, Flavián C, Guinalu M. New members' integration: key factor of success in online travel communities. *J Bus Res* 2013, 66:706–710.
122. Househ M, Borycki E, Kushniruk A. Empowering patients through social media: the benefits and challenges. *Health Inform J* 2014, 20:50–58.
123. Spence PR, Lachlan KA, Westerman D, Spates SA. Where the gates matter less: ethnicity and perceived source credibility in social media health messages. *Howard J Commun* 2013, 24:1–16.
124. Syed-Abdul S, Fernandez-Luque L, Jian W-S, Li Y-C, Crain S, Hsu M-H, Wang Y-C, Khandregzen D, Chuluunbaatar E, Nguyen PA, et al. Misleading health-related information promoted through video-based social media: anorexia on youtube. *J Med Internet Res* 2013, 15:e30.
125. Berry TR, Shields C. Source attribution and credibility of health and appearance exercise advertisements: relationship with implicit and explicit attitudes and intentions. *J Health Psychol* 2014, 19:242–252.
126. Dutta-Bergman MJ. The impact of completeness and web use motivation on the credibility of e-health information. *J Commun* 2004, 54:253–269.
127. Oh S, Yi YJ, Worrall A. Quality of health answers in social q&a. *Proc Am Soc Inf Sci Technol* 2012, 49:1–6.
128. Syn SY, Kim SU. The impact of source credibility on young adults' health information activities on facebook: preliminary findings. *Proc Am Soc Inf Sci Technol* 2013, 50:1–4.
129. Kwan MYW, Arbour-Nicitopoulos KP, Lowe D, Taman S, Faulkner GE. Student reception, sources, and believability of health-related information. *J Am Coll Health* 2010, 58:555–562.

130. Lederman R, Fan H, Smith S, Chang S. Who can you trust? Credibility assessment in online health forums. *Health Policy Technol* 2014, 3:13–25.

131. Festinger L. A theory of social comparison processes. *Hum Relat* 1954, 7:117–140.

132. McPherson M, Smith-Lovin L, Cook JM. Birds of a feather: homophily in social networks. *Annu Rev Sociol* 2001, 27:415–444.

133. Abbasi A, Fu T, Zeng D, Adjeroh D. Crawling credible online medical sentiments for social intelligence. In: *2013 International Conference on Social Computing* (*SocialCom*). IEEE; 2013, 254–263.

134. Pant G, Srinivasan P. Learning to crawl: comparing classification schemes. *ACM Trans Inf Syst* 2005, 23:430–462.

135. Weitzel L, de Oliveira JPM, Quaresma P. Measuring the reputation in user-generated-content systems based on health information. *Procedia Comput Sci* 2014, 29:364–378.

136. Freeman LC. Centrality in social networks conceptual clarification. *Social Netw* 1978, 1:215–239.

137. Mukherjee S, Weikum G, Danescu-Niculescu-Mizil C. People on drugs: credibility of user statements in health communities. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY: ACM; 2014, 65–74.

