# Protecting the Web from Misinformation

**Francesca Spezzano and Indhumathi Gurunathan**

**Abstract** Nowadays, a huge part of the information present on the Web is delivered through Social Media and User-Generated Content (UGC) platforms, such as Quora, Wikipedia, YouTube, Yelp, Slashdot.org, Stack Overflow, Amazon product reviews, and much more. Here, many users create, manipulate, and consume content every day. Thanks to the mechanism by which anyone can edit these platforms, its content grows and is kept constantly updated. However, malicious users can take advantage of this open editing mechanism to introduce misinformation on the Web.

In this chapter, we focus on Wikipedia, one of the main UCC platform and source of information for many, and study the problem of protecting Wikipedia articles from misinformation such as vandalism, libel, spam, etc. We address the problem from two perspectives: detecting malicious users to block such as spammers or vandals and detecting articles to protect, i.e., placing restrictions on the type of users that can edit an article. Our solution does not look at the content of the edits but leverages the users' editing behavior so that it generally results applicable to many languages. Our experimental results show that we are able to classify (1) article pages to protect with an accuracy greater than 92% across multiple languages and (2) spammers from benign users with 80.8% of accuracy and 0.88 mean average precision.

The chapter also defines different types of misinformation that exist on the Web and provides a survey of the methods proposed in the literature to prevent misinformation on Wikipedia and other platforms.

## 1 Introduction

Nowadays, a huge part of the information present on the Web is delivered through Social Media such as Twitter, Facebook, Instagram, etc., and User-Generated Content (UGC) platforms, such as Quora, Wikipedia, YouTube, Yelp, Slashdot.org,

F. Spezzano (✉) · I. Gurunathan
Computer Science Department, Boise State University, Boise, ID, USA
e-mail: francescaspezzano@boisestate.edu; indhumathigurunathan@u.boisestate.edu

© Springer Nature Switzerland AG 2020

M. A. Tayebi et al. (eds.), *Open Source Intelligence and Cyber Crime*, Lecture Notes in Social Networks, https://doi.org/10.1007/978-3-030-41251-7_1

Stack Overflow, Amazon product reviews, and many others. Here, users create, manipulate, and consume content every day. Thanks to the mechanism by which anyone can edit these platforms, its content grows and is kept constantly updated.

Unfortunately, Web features that allow for such openness have also made it increasingly easy to abuse this trust, and as people are generally awash in information, they can sometimes have difficulty discerning fake stories or images from truthful information. They may also lean too heavily on information providers or social media platforms such as Facebook to mediate even though such providers do not commonly validate sources. For example, most high school teens using Facebook do not validate news on this platform. The Web is open to anyone, and malicious users shielded by their anonymity threaten the safety, trustworthiness, and usefulness of the Web; numerous malicious actors potentially put other users at risk as they intentionally attempt to distort information, manipulate opinions and public response. Even worse, people can get paid to create fake news and spam reviews, influential bots can easily create it, and misinformation spreads so fast that is too hard to control. Impacts are already destabilizing the U.S. electoral system and affecting civil discourse, perception, and actions since what people read on the Web and events they think happened may be incorrect, and people may feel uncertain about their ability to trust it.

Misinformation can manifest in multiple forms such as vandalism, spam, rumors, hoaxes, fake news, clickbaits, fake product reviews, etc. In this chapter, we start by defining misinformation and describing different forms of misinformation that exist nowadays on the Web. Next, we focus on how to protect the Web from misinformation and provide a survey of the methods proposed in the literature to detect misinformation on social media and user-generated contributed platforms. Finally, we focus on Wikipedia, one of the main UGC platform and source of information for many, and study the problem of protecting Wikipedia articles from misinformation such as vandalism, libel, spam, etc. We address the problem from two perspectives: detecting malicious users to block such as spammers or vandals and detecting articles to protect, i.e., placing restrictions on the type of users that can edit an article. Our solution does not look at the content of the edits but leverages the users' editing behavior so that it generally results applicable to many languages. Our experimental results show that we are able to classify (1) article pages to protect with an accuracy greater than 92% across multiple languages and (2) spammers from benign users with 80.8% of accuracy and 0.88 mean average precision. Moreover, we discuss one of the main side effects of deploying anti-vandalism tools on Wikipedia, i.e. a low rate of newcomers retention, and an algorithm we proposed to early detect whether or not a user will become inactive and leave the community so that recovery actions can be performed on time to try to keep them contributing longer.

This chapter differs from the one by Wu et al. [1] because we focus more on the Wikipedia case study and how to protect this platform from misinformation, while Wu et al. mainly deal with rumors and fake news identification and intervention. Other related surveys are the one by Shu et al. [2] that focuses specifically on fake news, the work by Zubiaga [3] that deals with rumors, and the survey by Kumar and Shah [4] on fake news, fraudulent reviews, and hoaxes.

## 2 Misinformation on the Web

According to the Oxford dictionary, *misinformation* is “false or inaccurate information, especially that which is deliberately intended to deceive”. These days, the massive growth of the Web and social media has provided fertile ground to consume and quickly spread the misinformation without fact-checking. Misinformation can assume many different forms such as *vandalism, spam, rumors, hoaxes, counterfeit websites, fake product reviews, fake news, etc.*

Social media and user-generated content platforms like Wikipedia and Q&A websites are more likely affected by vandalism, spam, and abuse of the content. **Vandalism** is the action involving deliberate damage to others property, and Wikipedia defines vandalism on its platform as “the act of editing the project in a malicious manner that is intentionally disruptive” [5]. Beyond Wikipedia, other user-generated content platforms on the Internet got affected by vandalism. For example, editing/down-voting other users content in Q&A websites like Quora, Stack Overflow, Slashdot.org, etc. Vandalism can also happen on social media such as Facebook. For instance, the Martin Luther King, Jr.’s fan page was vandalized in Jan 2011 with racist images and messages.

**Spam** is, instead, a forced message or irrelevant content sent to a user who would not choose to receive it. For example, sending email to a bulk of users, flooding the websites with commercial ads, adding external link to the articles for promoting purposes, improper citations/references, spreading links created with the intent to harm, mislead or damage a user or stealing personal information, likejacking (tricking users to post a Facebook status update for a certain site without the user’s prior knowledge or intent), etc.

Wikipedia, like most forms of online social media, receives continuous spamming attempts every day. Since the majority of the pages are open for editing by any user, it inevitably happens that malicious users have the opportunity to post spam messages into any open page. These messages remain on the page until they are discovered and removed by another user. Specifically, Wikipedia recognizes three main types of spam, namely “advertisements masquerading as articles, external link spamming, and adding references with the aim of promoting the author or the work being referenced” [6].

User-generated content platforms define policies and methods to report vandalism, and spam and the moderation team took necessary steps like warning the user, blocking the user from editing, collapse the content if it is misinformation, block the question from visible to other users, or ban the user from writing/editing answers, etc. These sites are organized and maintained by the users and built as a community. So the users have responsibilities to avoid vandalism and make it as a knowledgeable resource to others. For example, the Wikipedia community adopts several mechanisms to prevent damage or disruption to the encyclopedia by malicious users and ensure content quality. These include administrators to ban or block users or IP addresses from editing any Wikipedia page either for a finite amount of time or indefinitely, protecting pages from editing, or detecting damaging content to be reverted through dedicated bots [7, 8], monitoring recent changes, or having watch-lists.

Slashdot gives moderator access to its users to do jury duty by reading comments and flag it with appropriate tags like Offtopic, Flamebait, Troll, Redundant, etc. Slashdot editors also act as moderators to downvote abusive comments. In addition to that, there is an "Anti" symbol present for each comment to report spam, racist ranting comments, etc. Malicious users can also act protected by anonymity. In Quora, if an anonymous user vandalizes the content, then a warning message is sent to that user's inbox without revealing the identity. If the particular anonymous user keeps on abusing the content, then Quora moderator revokes the anonymity privileges of that user.

Online reviews are not free from misinformation either. For instance, on Amazon or Yelp, it is frequent to have spam paid reviewers writing **fraudulent reviews** (or opinion spam) to promote or demote products or businesses. Online reviews help customers to make decisions on buying the products or services, but when the reviews are manipulated, it will impact both customers and business [9]. Fraudulent reviewers post either positive review to promote the business and receive something as compensation, or they write negative reviews and get paid by the competitors to create damage to the business. There are some online tools like `fakespot.com` and `reviewmeta.com` that analyze the reviews and helps to make decisions. But in general, consumers have to use some common sense to not fall for fraudulent reviews and do some analysis to differentiate the fake and real reviews. Simple steps like verifying the profile picture of the reviewer, how many other reviews they wrote, paying attention to the details, checking the timestamp, etc., will help to identify fraudulent reviews.

Companies also take some actions against fraudulent reviews. Amazon sued over 1000 people who posted fraudulent reviews for cash. It is also suspending the sellers and shut-downing their accounts if they buy fraudulent reviews for their products. They rank the reviews and access the buyer database to mark the review as "Verified Purchase" meaning that the customer who wrote the review also purchased the item at Amazon.com. Yelp has an automated filtering software that is continuously running to examine each review recommend only useful and reliable reviews to its consumers. Yelp also leverages the crowd (consumer community) to flag suspicious reviews and takes legal action against the users who are buying or selling reviews.

**Fake news** is low-quality news that is created to spread misinformation and misleading readers. The consumption of news from social media is highly increased nowadays so as spreading of fake news. According to the Pew research center [10], 64% of Americans believe that fake news causes confusion about the basic facts of current events. A recent study conducted on Twitter [11] revealed that fake news spread significantly more than real ones, in a deeper and faster manner and that the users responsible for their spread had, on average, significantly fewer followers, followed significantly fewer people, were significantly less active on Twitter. Moreover, bots are equally responsible for spreading real and fake news, and then the considerable spread of fake news on Twitter is caused by human activity.

Fact-checking the news is important before spreading it on the Web. There are a number of news verifying websites that can help consumers to identify fake news by making more warranted conclusions in a fraction of the time. Some examples of fact-checkers are `FactCheck.org`, `PolitiFact.com`, `snopes.com`, or `mediabiasfactcheck.com`.

Beyond fact-checking, consumers should also be responsible for [12]:

1. *Read more than the headline*—Often fake news headlines are sensational to provoke readers emotions that help the spread of fake news when readers share or post without reading the full story.
2. *Check the author*—The author page of the news website provides details about the authors who wrote the news articles. The credibility of the author helps to measure the credibility of the news.
3. *Consider the source*—Before sharing the news on social media, one has to ensure the source of the articles, verify the quotes that the author used in the article. Also, a fake news site often has strange URL's.
4. *Check the date*—Fake news sometimes provides links to previously happened incidents to the current events. So, one needs to check the date of the claim.
5. *Check the bias*—If the reader has opinion or beliefs to one party, then they tend to believe biased articles. According to a study done by Allcott and Gentzkow [13], the right-biased articles are more likely to be considered as fake news.

Moreover, one of the most promising approaches to combat fake news is promoting *news literacy*. Policymakers, educators, librarians, and educational institutions can all help in educating the public—especially younger generations—across all platforms and mediums [14].

**Clickbait** is a form of link-spam leading to fake content (either news or image). It is a link with a catchy headline that tempts users to click on the link, but it leads to the content entirely unrelated to the headline or less important information. Clickbait works by increasing the curiosity of the user to click the link or image. The purpose of a clickbait is to increase the page views which in turn increase the revenue through ad sense. But when it is used correctly, the publisher can get the readers attention, if not the user might leave the page immediately. Publishers employ various cognitive tricks to make the readers click the links. They write headlines to grab the attention of the readers by provoking their emotions like anger, anxiety, humor, excitement, inspiration, surprise. Another way is by increasing the curiosity of the readers by presenting them with something they know a little bit but not many details about the topic. For example, headlines like "You won't believe what happens next?" provoke the curiosity of the readers and make them click.

**Rumors** are pieces of information whose veracity is unverifiable and spreads very easily. Their source is unknown, so most of the time the rumors are destructive and misleading. Rumors start as something true and get exaggerated to the point that it is hard to prove. They are often associated with breaking news stories [15]. Kwon et al. [16] report on many interesting findings on rumor spreading dynamics such as (1) a rumor flows from low-degree users to high-degree users, (2) a rumor rarely initiate a conversation and people use speculative words to express doubts about their validity when discussing rumors, and that rumors do not necessarily contain different sentiments than non-rumors. Friggeri et al. [17] analyzed the propagation of known rumors from `Snopes.com` in Facebook and their evolution over time. They found that rumors run deeper in the social network than reshare cascades in general and that when a comment refers to a rumor and contains a link to a Snopes article, then the likelihood that a reshare of a rumor will be deleted increases. Unlike rumors, **hoaxes** consist of false information pretending to be true information and often intended as a joke. Kumar et al. [18] show that 90% of hoaxes articles in Wikipedia are identified in 1 h after their approval, while 1% of hoaxes survive for over 1 year.

Misinformation is also spread through **counterfeit websites** that disguise as legitimate sites. For instance, `ABCnews.com.co` and `Bloomberg.ma` are examples of fake websites. They create more impact and cause severe damage when these sites happen to be subject specific to medical, business, etc.

Also, online videos can contain misinformation. For instance, Youtube videos can have clickbaiting titles, spam in the description, inappropriate or not relevant tags to the videos, etc. [19]. This metadata is used to search and retrieve the video and misinformation in the title or the tags lead to increase the video's views and, consequently, the user' monetization. Sometimes online videos are entirely fake and can be automatically generated via machine learning techniques [20]. As compared to recorded videos, computer-generated ones lack the imperfections, a feature that is hard to incorporate in a machine-learning based algorithm to detect fake videos [21].

## 3 Detecting Misinformation on the Web

To protect the Web from misinformation, researchers focused on detecting misbehavior, i.e., malicious users such as vandals, spammers, fraudulent reviewers, rumors and fake news spreaders that are responsible for creating and sharing misinformation, or detecting whether or not a given piece of information is false.

In the following, we survey the main methods proposed in the literature to detect either the piece of misinformation or the user causing it. Table 1 summarizes all the related work grouped by misinformation type.

### *3.1 Vandalism*

Plenty of work has been done on detecting vandalism, especially on Wikipedia. One of the first works is the one by Potthast et al. [22] that uses feature extraction (including some linguistic features) and machine learning and validate them on

**Table 1** Related work in detecting misinformation by type

| Types of misinformation | Related work |
|---|---|
| Vandalism | [22–30] |
| Spam | [31–44] |
| Fraudulent reviews | [45–56] |
| Fake news | [2, 4, 57–67] |
| Clickbaits | [68–71] |
| Rumors | [3, 72–80] |
| Hoaxes | [18, 81] |

the PAN-WVC-10 corpus: a set of 32K edits annotated by humans on Amazon Mechanical Turk [23]. Adler et al. [24] combined and tested a variety of proposed approaches for vandalism detection including natural language, metadata [25], and reputation features [26]. Kiesel et al. [27] performed a spatiotemporal analysis of Wikipedia vandalism revealing that vandalism strongly depends on time, country, culture, and language. Beyond Wikipedia, vandalism detection has also been addressed in other platforms such as Wikidata [28] (the Wikimedia knowledge base) and OpenStreetMaps [29].

Currently, ClueBot NG [7] and STiki [8] are the state-of-the-art tools used by Wikipedia to detect vandalism. ClueBot NG is a bot based on an artificial neural network which scores edits and reverts the worst-scoring edits. STiki is an intelligent routing tool which suggests potential vandalism to humans for definitive classification. It works by scoring edits by metadata and reverts and computing a reputation score for each user. Recently, Wikimedia Foundation launched a new machine learning-based service, called Objective Revision Evaluation Service (ORES) [82] which measures the level of general damage each edit causes. More specifically, given an edit, ORES provides three probabilities predicting (1) whether or not it causes damage, (2) if it was saved in good-faith, and (3) if the edit will eventually be reverted. These scores are available through the ORES public API [83].

In our previous work [30], we addressed the problem of vandalism in Wikipedia from a different perspective. We studied for the *first* time the problem of detecting vandal users and proposed VEWS, an early warning system to detect vandals before other Wikipedia bots.[1] Our system leverages differences in the editing behavior of vandals vs. benign users and detect vandals with an accuracy of over 85% and outperforms both ClueBot NG and STiki. Moreover, as an early warning system, VEWS detects, on average, vandals 2.39 edits before ClueBot NG. The combination of VEWS and Cluebot NG results in a fully automated system that does not leverage any human input (e.g., edit reversion) and further increases the performances.

Another mechanism used by Wikipedia to protect against content damage is *page protection*, i.e., placing restrictions on the type of user that can edit the page. To the best of our knowledge, little research has been done on the topic of page protection

[1] Dataset and code are available at http://www.cs.umd.edu/~vs/vews/.

in Wikipedia. Hill and Shaw [84] studied the impact of page protection on user patterns of editing. They also created a dataset (they admit it may not be complete) of protected pages to perform their analysis. There are not currently bots on Wikipedia that can search for pages that may need to be protected. Wikimedia does have a script [85] available in which administrative users can protect a set of pages all at once. However, this program requires that the user supply the pages or the category of pages to be protected and is only intended for protecting a large group of pages at once. There are some bots on Wikipedia that can help with some of the wiki-work that goes along with protecting or removing page protection. This includes adding or removing a template to a page that is marked as protected or no longer marked as protected. These bots can automatically update templates if page protection has expired.

## *3.2 Spam*

Regarding spam detection, various efforts have been made to detect spam users on social networks, mainly by studying their behavior after collecting their profiles through deployed social honeypots [31, 32]. Generally, social networks properties [33, 34], posts content [35, 36], and sentiment analysis [37] have been used to train classifiers for spam users detection.

Regarding spam detection in posted content specifically, researchers mainly concentrated on the problem of predicting whether a link contained in an edit is spam or not. URLs have been analyzed by using blacklists, extracting lexical features and redirecting patterns from them, considering metadata or the content of the landing page, or examining the behavior of who is posting the URL and who is clicking on it [38–41]. Another big challenge is to recognize a short URL as spam or not [42].

Link-spamming has also been studied in the context of Wikipedia. West et al. [43] created the first Wikipedia link-spam corpus, identified Wikipedia's link spam vulnerabilities, and proposed mitigation strategies based on explicit edit approval, refinement of account privileges, and detecting potential spam edits through a machine learning framework. The latter strategy, described by the same authors in [44], relies on features based on (1) article metadata and link/URL properties, (2) HTML landing site analysis, and (3) third-party services used to discern spam landing sites. This tool was implemented as part of STiki (a tool suggesting potential vandalism) and has been used on Wikipedia since 2011. Nowadays, this STiki component is inactive due to a monetary cost for third-party services.

## *3.3 Rumors and Hoaxes*

The majority of the work focused on studying rumors and hoaxes characteristics, and very little work has been done on automatic classification [3, 72, 73]. Qazvinian et al. [74] addressed the problem of rumor detection in Twitter via temporal, content-based and network-based features and additional features extracted from hashtags and URLs present in the tweet. These features are also effective in identifying disinformers, e.g., users who endorse a rumor and further help it to spread. Zubiaga et al. [75] identify whether or not a tweet is a rumor by using the context of from earlier posts associated with a particular event. Wu et al. [76] focused on early detection of emerging rumors by exploiting knowledge learned from historical data. More work has been done for rumor or meme source identification in social networks by defining ad-hoc centrality measures, e.g., rumor centrality, and study rumor propagation via diffusion models, e.g., the SIR model [77–80].

Kumar et al. [18] proposed an approach to detect hoaxes according to article structure and content, hyperlink network properties, and hoaxes' creator reputation. Tacchini et al. [81] proposed a technique to classify Facebook posts as hoaxes or non-hoaxes on the basis of the users who "liked" them.

## *3.4 Fraudulent Reviews*

A Fraudulent review (or deceptive opinion spam) is a review with fictitious opinions which are deliberately written to sound authentic. There are many characteristics that are often hallmarks of fraudulent reviews:

1. *There is no information about the reviewer.* Users who only post a small number of reviews or have no profile information or social connections are more likely to be fibbing.
2. *The opinions are all-or-nothing.* Fabricated reviews tend to be more extreme (all 5 stars or all one star).
3. *Several are posted at once.* Suddenly a product or company with no reviews or one every few months will have five in a row all mentioning something similar, from the same day which indicates that a company paid for a batch of reviews.
4. *They use smaller words.* Scientists say it takes more brainpower to tell a lie than a truth; when we're telling a lie, our vocabulary tends to suffer because we're already expending mental energy on the fabrication. As a result, fraudulent reviews are characterized by shorter words according to research.
5. *They are very short.* Since fraudulent review mills may only pay a few dollars (or less) per review, there's an incentive for a writer to dash them off quickly.

Jindal and Liu [45] first studied deceptive opinion spam problem and trained machine learning-based models using features based on the opinion content, user, and the product itself. Ott et al. [46] created a benchmark dataset by collecting real reviews from TripAdvisor and employing Amazon Mechanical Turk workers to write fraudulent reviews. They got 90% accuracy in detecting fraudulent reviews on their dataset by using psycholinguistic-based features and text-based features (bigrams). However, Mukherjee et al. [47] found out that the method proposed by Ott et al. is not enough to have good performances on a larger and more realistic dataset extracted from Yelp, but behavioral features on the user who wrote the review performed very well (86% accuracy). They also reported that the word distribution between fake and real reviews is very different in the dataset by Ott et al., while this is not true in their more realistic Yelp dataset.

Since then, researchers started focusing more on the problem of detecting opinion spammers (or fraudulent reviewers), rather than fraudulent reviews. Fei et al. [48] discovered that a large number of opinions made use of a sudden burst either caused by the sudden popularity of the product or by a sudden invasion of a large number of fake opinions including some of the features of real users. They used this finding to design an algorithm that applies loopy belief propagation on a network of reviewers appearing in different bursts to detect opinion spammers. Several other works considered features extracted from reviewer behavior as well [49–51].

Rayana and Akoglu [52] also applied loopy belief propagation on a user–product bipartite network with signed edges (positive or negative reviews) and considered metadata (text, timestamp, rating) to assign prior probabilities of users being spammers, reviews being fake, and products being targeted by spammers. Wang et al. [53] proposed three measures (the trustworthiness of the user, the honesty of the review, and the reliability of the store) to be computed on the user-product-store network. Hooi et al. [54] proposed the BirdNest algorithm that detects opinion spammers according to the fact that (1) fraudulent reviews occur in short bursts of time and (2) fraudulent user accounts have skewed rating distributions. Kumar et al. [55] bridged network data and behavioral data to define measures for computing the fairness (or trustworthiness) of the reviewer, the goodness (or quality) of the product, and the reliability of the review. Several other works have been proposed to detect a group of opinion spammers. For instance, CopyCatch [56] leveraged the lockstep behavior, i.e., groups of users acting together, generally liking the same pages at around the same time.

## *3.5 Fake News*

To identify fake news, the majority of the approaches proposed in the literature have focused on machine learning-based approaches working with features extracted from news content and social context [2].

*News content-based features* include both linguistic features extracted from the text of the news, metadata-based features such as news source (author and/or publisher), headlines, etc., and visual-based features extracted from images and videos associated with the news. For instance, Seyedmehdi and Papalexakis [57] proposed a solution based on extracting latent features from news article text via tensor decomposition to categorize fake news as extreme bias, conspiracy theory, satire, junk science, hate group, or state news. Potthast et al. [58] used the writing style of the articles to identify extremely biased news from the neutral one by using the techniques called unmasking. This model used the news domain specific style features like ratios of quoted words, external links, the average length of the paragraph, etc. Horne and Adali [59] considered both news body and headline for determining the validity of news. They found out that fake and real news have drastically different headlines as they were able to obtain a 0.71 accuracy when considering the number of nouns, lexical redundancy (TTR), word count, and the number of quotes. Further, the study found that fake titles contain different sorts of words (stop words, extremely positive words, and slang among others) than titles of real news articles. Pérez-Rosas et al. [60] analyzed the news body content only and achieved an accuracy up to 0.76 in detecting fake news. They also tested cross domain classification obtaining poor performances by training in one dataset and testing in the other one. Jin et al. [61] used only visual and statistical features extracted from news images for microblogs news verification.

*Social context-based features* consider (1) the profile and characteristics of users creating and spreading the news (e.g., number of followers/followees, number of posts, credibility and the reliability of the user) also averaged among all the users related to particular news, (2) users' opinion and reactions towards social media posts (post can potentially contain fake news), (3) various type of networks such as friendship networks, co-occurrence networks (network formed based on the number of posts the user write related to the news), or diffusion network where edges between users represent information dissemination paths among them.

Kim et al. [62] propose methods to not only detect the fake news but also to prevent the spread of fake news by making the user flag fake news and used reliable third-parties to fact check the news content. They developed an online algorithm for this purpose, so it works at the time of user spreading the fake news thus preventing it from spreading. Jin et al. [63] developed a method for detecting fake news by using the users' viewpoints to find relationships such as support or oppose and by building a credibility propagation network by using these relationships. Wu and Liu [64] used the way news spread through the social network to find the fake news. They used graph mining method to analyze the social network and recurrent neural networks to represent and classify propagation pathways of a message.

Finally, *hybrid methods* combine the two previous approaches. For instance, Ruchansky et al. [65] used temporal behavior of users and their response and the text content of the news to detect the fake news. They proposed the CSI model (Capture, Score, and Integrate) to classify the news article. Fairbanks et al. [66] show that a content-based model can identify publisher political bias while a structural analysis of web links is enough to detect whether the news is credible or not. Shu et al. [67] exploited both fake news content and the relationship among publishers, news pieces, and users to detect fake news.

Regarding *clickbait detection* specifically, Chakraborty et al. [68] build personalized automatic blocker for clickbait headlines by using a rich set of features that use sentence structure, word patterns, N-gram features, and clickbait language. Their browser extension 'Stop-Clickbait' warns users for potential clickbaited headlines. Potthast et al. [69] used Twitter datasets to identify messages in social media that lead to clickbait. They gathered tweets from various publishers and constructed features based on teaser message, linked web page, and meta information. Anand et al. [70] used three variants of bidirectional RNN models (LSTM, GPU, and standard RNNs) for detecting clickbait headlines They used two different word embedding techniques such as distributed word embeddings and character-level word embeddings. Chen et al. [71] examined a hybrid approach for clickbait detection by using text-based and non-text based clickbaiting cues. While textual cues use text-based semantic and syntactical analysis, non-textual cues relate to image and user behavior analysis.

## 4 Case Study: Protecting Wikipedia Content Quality

Wikipedia is the world's biggest free encyclopedia read by many users every day. Thanks to the mechanism by which anyone can edit, its content is expanded and kept constantly updated. However, malicious users can take advantage of this open editing mechanism to seriously compromise the quality of Wikipedia articles. As we have seen in Sect. 2, the main form of content damaging in Wikipedia is vandalism, but other types of damaging edits are also common such as page spamming [6] and dissemination of false information, e.g., through hoax articles [86].

In this section, we discuss our research effort to ensure the content integrity of Wikipedia. W start by introducing the DePP system, which is the state-of-the-art tool for detecting article pages to protect [87, 88] in Wikipedia. Page protection is a mechanism used by Wikipedia to place restrictions on the type of users that can make edits to prevent vandalism, libel, or edit wars. Our DePP system achieves an accuracy of 92.1% across multiple languages and significantly improves over baselines.

Then, we present our work on spam users identification [89]. We formulate the problem as a binary classification task and propose a new system, called WiSDe, based on a set of features representing user editing behavior to separate spam users from benign ones. Our results show that WiSDe reaches 80.8% classification accuracy and 0.88 mean average precision and beat ORES, the most recent tool developed by Wikimedia to assign damaging scores to edits.

Finally, we discuss our related work [90, 91] on detecting editors who will stop contributing to the encyclopedia. In fact, one of the main problems of fighting vandalism on Wikipedia is that newcomers are considered suspicious from veteran users who often delete their contributions causing the non-integration of newcomers in the community. We think that the early prediction of inactive users is useful for Wikipedia administrators or other users to perform recovering actions in time to avoid the loss of contributors. This section does not provide new results but collects the ones presented in our prior publications.

### *4.1 Detecting Pages to Protect*

The first problem we address consists of deciding whether or not Wikipedia administrators should protect a page. *Page protection* consists of placing restrictions on the type of users that can edit a Wikipedia page. Examples of protected pages on English Wikipedia are *Drug* and *Biology*. Users can recognize this kind of pages by the image of a lock in the upper right-hand corner of the page. Common motivations that an administrative user may have in protecting a page include (1) consistent vandalism or libel from one or more users, and (2) avoiding edit wars [92]. An edit war is when two users cannot agree on the content of an article and one user repeatedly reverts the other's edits.

There are different levels of page protection for which different levels of users can make edits (or, in general, perform actions on the page): fully protected pages can be edited (or moved) only by administrators, semi-protected pages can be modified only by autoconfirmed users, while move protection does not allow pages to be moved to a new title, except by an administrator. Page protections can also be set for different amounts of time, including 24 or 36 h, or indefinitely.

Currently, the English Wikipedia contains over five million pages. Only a small percentage of those pages are currently protected less than 0.2%. However, around 17 pages become protected every day (according to the number of protected pages from May 6 through Aug 6, 2016). This ratio shows how it is difficult for administrative users to monitor overall Wikipedia pages to determine if any need to be protected. Users can request pages to be protected or unprotected, but an administrative user would have to analyze the page to determine if it should be protected, what level of protection to give, and for how long the protection should last, if not indefinitely. All this work is currently *manually* done by administrators.

To overcome this problem, we propose `DePP`, the *first* automated tool to detect pages to protect in Wikipedia. `DePP` is a machine learning-based tool that works with two novel set of features based on (1) *users page revision behavior* and (2) *page categories*. More specifically, the first group of features includes the following six base features:

E1 *Total average time between revisions*: pages that have very few edits over a long period of time are less likely to become protected (as their content is more stable) than pages with many edits that happen with little time between them.

E2 *Total number of users making five or more revisions*: this feature counts the number of users who make more than five edits to a page.

E3 *Total average number of revisions per user*: if there are many users making a few changes to a page, it is less likely to become protected than if a few users are making a lot of changes to a page.

E4 *Total number of revisions by non-registered users*: this feature measures the number of changes made to a page from non-registered users. If a user has not spent the time to set up an account, it is less likely that they are a proficient user and more likely to be a spammer or vandal. Therefore, the more non-registered users that are editing a page, the more likely it is that the page may need to be protected.

E5 *Total number of revisions made from mobile device*: similar to feature E4, this feature looks at the number of revisions that are tagged as coming from a mobile device. This is a useful feature because users making changes from a mobile device are not likely to be sitting down to spend time making revisions to a page that would add a lot of value. It is possible that a user making a change from a mobile device is only adding non-useful information, vandalizing a page, or reverting vandalism that needs to be removed immediately.

E6 *Total average size of revisions*: it is possible that users vandalizing a page, or adding non-useful information would make an edit that is smaller in size. This is opposed to a proficient user who may be adding a large amount of new content to a page. For this reason, we measure the average size of an edit. Small edits to a page may lead to a page becoming protected more than large edits would.

In addition to the above base features, we also include an additional set of features taking into account the page editing pattern over time. We define these features by leveraging the features E1–E6 as follows. For each page, we consider the edits made in the latest 10 weeks and we split this time interval into time frames of 2 weeks (last 2 weeks, second last 2 weeks, etc.). Then, we compute the base features E1–E6 within each time frame and compute the standard deviation of each base features in the 10 weeks time interval. This produces six new features whose idea is to measure how much features E1–E6 are stable over time. For instance, for normal pages with solid content, we may observe fewer edits of smaller size representing small changes in the page, corresponding to a low standard deviation for features E1, E3, and E6. On the other hand, a higher standard deviation of the base features can describe a situation where the content of the page was initially stable, but suddenly we observe a lot of edits from many users, which may indicate the page is under a vandalism attack and may need protection.

The second group of features use information about page categories and includes:

NC *Number of categories a page is marked under*;

PC *Probability of protecting the page given its categories*: given all the pages in the training set $T$ and a page category $c$, we compute the probability $\texttt{pr}(c)$ that pages in category $c$ are protected as the percentage of pages in $T$ having category $c$ that are protected. Then, given a page $p$ having categories $c_1, \ldots, c_n$, we compute this feature as the probability that the page is in at least one category whose pages have a high probability to be protected as

$$PC(p) = 1 - \prod_{i=1}^{n}(1 - \texttt{pr}(c_i)).$$

We also define another group of features that shows how much features E1–E6 vary for a page $p$ w.r.t. the average of these values among all the pages in the same categories as $p$. Specifically, given the set of pages in the training set $T$, we computed the set $C$ of the top-$k$ most frequent categories. Additionally, for each category $c \in C$, we averaged the features E1–E6 among all the pages (denoted by $T_c$) having the category $c$ in the training set. Then, for each page $p$ we computed the following set of features, one for each feature E$i$ ($1 \leq i \leq 6$) and for each category $c \in C$ as follows:

$$C(Ei, c) = \begin{cases} Ei(p) - \mathtt{avg}_{p' \in T_c}(Ei(p')) & \text{if } p \text{ is in category } c \\ 0 & \text{otherwise} \end{cases}$$

where $Ei(p)$ is the value of the feature E$i$ for the page $p$. The aim of this group of features is to understand if a page is anomalous w.r.t. other pages in the same category. All the features that we propose are language independent as they do not consider page content. As a consequence, DePP is general and able to work on any version of Wikipedia.

To test our DePP system, we built four balanced datasets, one for each of the following Wikipedia versions: English, German, French, and Italian. Each dataset contains all edit protected articles until to Oct. 12, 2016, an almost equal number of randomly selected unprotected pages, and up to the last 500 most recent revisions for each selected page. The sizes of these datasets[2] are reported in Table 2. For protected pages, we only gathered the revisions up until the most recent protection. If there was more than one recent protection, we gathered the revision information between the two protections. This allowed us to focus on the revisions leading up to the most recent page protection. Revision information that we collected included the user who made the revision, the timestamp of the revision, the size of the revision, the categories of the page, and any comments, tags or flags associated with the revision.

The DePP accuracy in the prediction task on 10-fold cross validation is reported for random forest (the best performing algorithm as compared to Logistic Regression, SVM, and K-Nearest Neighbor) in Table 3. As we can see, DePP can classify pages to protect from pages that do not need protection with an accuracy greater

**Table 2** English, German, French, and Italian Wikipedia datasets used in to test our DePP system

| | English | German | French | Italian |
|---|---|---|---|---|
| Protected pages | 7968 | 1722 | 524 | 171 |
| Unprotected pages | 7889 | 1706 | 512 | 168 |
| Number of edits | 2.2M | 311K | 106K | 29K |

**Table 3** DePP accuracy results and comparison with baselines

| | English | German | French | Italian |
|---|---|---|---|---|
| B1+B2+B3 | 78% | 48% | 77% | 43% |
| **DePP** | **95%** | **93%** | **93%** | **91%** |

Everything is computed with random forest (best classifier)
Best scores are highlighted in bold

[2] Datasets available at http://bit.ly/wiki_depp.

than 91% in all the four languages. As no automated tool detecting which page to protect exists in Wikipedia, we defined some baselines to compare our results. One of the main reasons for protecting a page on Wikipedia is to stop edit wars, vandalism or libel from happening or continuing to happen on a page. Thus, we used the following baselines:

B1 *Number of revisions tagged as "Possible libel or vandalism"*: These tags are added automatically without human interference by checking for certain words that might be likely to be vandalism. If a match is found, the tag is added.

B2 *Number of revisions that Wikipedia bots or tools reverted as possible vandalism*: number of reverted edits in the page made by each one of these tools. We considered Cluebot NG and STiki for English Wikipedia and Salebot for French Wikipedia. We did not find any bot fighting vandalism for German or Italian Wikipedia.

B3 *Number of edit wars between two users in the page*: Edit warring occurs when two users do not agree on the content of a page or revision. Therefore, we count the number of edit wars within the revision history of a page as another baseline. In some Wikipedia languages, e.g., German and Italian, there is an explicit tag denoting edit wars. For English and French Wikipedia, we define an edit war as one user making a revision to a page, followed by another user reverting that revision, and this pattern happens 2 or 3 consecutive times.

As we can see in Table 3, `DePP` significantly beats the combination of all the three baselines across all the languages. By analyzing the most important features, we found that features E1 (total average time between revisions) and PC (probability of protecting the page given its categories) consistently appear within the top-15 features in all the four languages considered. Wikipedia editors spend less time in revising pages that end up being protected. For instance, in English Wikipedia the mean average time between revisions in 5.8 days for protected pages and 2.9 months for unprotected ones. Also, a protected page is more likely to be in categories that have other protected pages than an unprotected page (a probability of 0.84 on average vs. 0.52 in English Wikipedia).

In the real-world scenario, we have more unprotected pages than unprotected ones.[3] Thus, we performed an experiment where we created an unbalanced setting by randomly selecting pages at a ratio of 10% protected and 90% unprotected (due to the size of the data we have, we could not reduce this ratio further). Then, we performed 10-fold cross-validation and measured the performance by using the area under the ROC curve (AUROC). We used class weighting to deal with class imbalance. Due to the randomness introduced, we repeated each experiment 10 times and averaged the results. Results are reported in Table 4. We observe that AUROC values are pretty high across all the dataset considered and outperforms

[3] Currently, there is a 0.16% of protected pages in English Wikipedia, 0,09% in German, 0.04% in French, and 0.015% in Italian.

**Table 4** DePP AUROC results and comparison with baselines in the unbalanced setting (10% protected pages, 90% unprotected)

| | English | German | French | Italian |
|---|---|---|---|---|
| B1+B2+B3 | 0.77 | 0.50 | 0.80 | 0.50 |
| **DePP** | **0.97** | **0.97** | **0.97** | **0.96** |

Everything is computed with random forest (best classifier)
Best scores are highlighted in bold

the baselines. In comparison, AUROC values for the balanced setting are 0.98 for English Wikipedia, 0.98 for German, 0.97 for French, and 0.93 for Italian. Thus, performance does not drop when considering a more real-world unbalanced scenario. Moreover, as shown in [93], AUROC values do not change with changes in the test distribution, thus the above AUROC values for the balanced setting are generalizable to an unbalanced one. Random forest results the best classifier in both the balanced and unbalanced setting.

## *4.2 Spam Users Identification*

Another problem that compromises the content quality of Wikipedia articles is spamming. Currently, no specific tool is available on Wikipedia to identify neither spam edits or spam users. Tools like Cluebot NG and STiki are tailored toward vandalism detection, while ORES is designed to detect damaging edits in general. As in the case of page protection, the majority of the work to protect Wikipedia from spammers is done *manually* by Wikipedia users (patrollers, watchlisters, and readers) who monitor recent changes in the encyclopedia and, eventually, report suspicious spam users to administrators for definitive account blocking. To fight spammers on Wikipedia, we study the problem of identifying spam users from good ones [89]. Our work is closer in spirit to [30] as the aim is to classify users by using their editing behavior instead of classifying a single edit as vandalism [7, 8], spam [44] or generally damaging [82].

We propose a system, called WiSDe (Wikipedia Spammer Detector), that uses a machine learning-based framework with a set of features which are based on research that has been done regarding typical behaviors exhibited by spammers: similarity in edit size and links used in revisions, similar time-sensitive behavior in edits, social involvement of a user in the community through contribution to Wikipedia's talk page system, and chosen username. We did not consider any feature related to edit content so that our system would be language independent and capable of working for all Wikipedia versions. Also, the duration of a user's edit history, from the first edit to her most recent edit, is not taken into account as this feature is biased towards spammers who are short-lived due to being blocked by administrators. Finally, we do not rely on third-party services, so there is no overhead cost as in [44].

The list of features we considered to build `WiSDe` are as follows:

**User's Edit Size Based Features**

S1 *Average size of edits*—since spammers in Wikipedia are primarily trying to promote themselves (or some organization) and/or attract users to click on various links, the sizes of spammers' edits are likely to exhibit some similarity when compared to that of benign users.

S2 *Standard deviation of edit sizes*—since many spammers make revisions with similar content, the variation in a user's edit sizes is likely not to be very large when compared to benign users.

S3 *Variance significance*—since variance in a spam user's edits can change based on a user's average edit size, normalizing a user's standard deviation of edit sizes by their average edit size may balance any difference found by considering the standard deviation alone.

**Editing Time Behavior Based Features**

S4 *Average time between edits*—spammers across other social media tend to perform edits in batches and in relatively rapid succession, while benign Wikipedia users dedicate more time in curating the article content and then make edits more slowly than spammers.

S5 *Standard deviation of time between edits*—the consistency in timing of spammers' edits tends to be somewhat mechanical, while benign users tend to edit more sporadically.

**Links in Edit Based Features**

S6 *Unique link ratio*—since spammers often post the same links in multiple edits, a measure of how unique any links that a user posts may be very useful in helping to determine which users are spammers. This measure is calculated for any user that has posted a minimum of two links in all of their edits, and it is the ratio of unique links posted by a user to the total number of links posted by the user (considering only the domain of the links)

S7 *Link ratio in edits*—since spammers on Wikipedia are known to post links in an effort to attract traffic to other sites the number of edits that a user makes which contain links is likely a useful measure in determining spammers from benign users.

**Talk Page Edit Ratio** Since talk pages do not face the public and are only presented to a user that specifically clicks on one, spammers are less likely to get very many views on these pages, and, therefore are much less likely to make edits to talk pages. Because of this, the ratio of talk pages edited by a user that correspond with the main article pages that a user edits is considered a possible good indicator of whether a user is a spammer or not. We denote this feature by S8.

**Username Based Features** Zafarani and Liu [94] showed that aspects of users' usernames themselves contain information that is useful in detecting malicious users. Thus, in addition to the features based on users' edit behaviors, we also considered four additional features related to the user's username itself. These four features are: the *number of digits in a username* (S9), the *ratio of digits in a username* (S10), the *number of leading digits in a username* (S11), and the *unique character ratio in a username* (S12).

To test our WiSDe system, we built a new dataset[4] containing 4.2K (half spammer and half benign) users and 75.6K edits as follows. We collected all Wikipedia users (up to Nov. 17, 2016) who were blocked for spamming from two lists maintained on Wikipedia: "Wikipedians who are indefinitely blocked for spamming" [95] and "Wikipedians who are indefinitely blocked for link spamming" [96]. The first list contains all spam users blocked before Mar 12, 2009, while the second one includes all link-spammers after Mar 12, 2009, to today. We gathered a total of 2087 spam users (we only included users who did at least one edit) between the two lists considered.

In order to create a balanced dataset of spam/benign users, we randomly select a sample of benign Wikipedia users of roughly the same size as the spammer user set (2119 users). To ensure these were genuine users, we cross-checked their usernames against the entire list of blocked users provided by Wikipedia [97]. This list contains all users in Wikipedia who have been blocked for any reason, spammers included. For each user in our dataset, we collected up to their 500 most recent edits. For each edit, we gathered the following information: edit content, time-stamp, whether or not the edit is done on a Talk page, and the damaging score provided by ORES.

We run 10-fold cross-validation on several machine learning algorithms, namely SVM, Logistic Regression, K-Nearest Neighbor, Random Forest, and XGBoost, to test the performances of our features. Experimental results are shown in Table 5 for the best performing algorithm (XGBoost). Here we can see that WiSDe is able to classify spammers from benign users with 80.8% of accuracy and it is a valuable tool in suggesting potential spammers to Wikipedia administrators for further investigation as proved by a mean average precision of 0.88.

Feature importance analysis revealed that the top three most important features for spammers identifications are: *Link ratio in edits*, *Average size of edits*, and *Standard deviation of time between edits*. As expected, spammers use more links in their edits. The average value of this feature is 0.49 for spammers and 0.251 for benign users. Also, benign users put more diverse links in their revisions than spammers (0.64 vs. 0.44 on average). We also have that spammer's edit size is

**Table 5** WiSDe spammers identification accuracy and Mean Average Precision (MAP) results in comparison with ORES

| | Accuracy | MAP |
|---|---|---|
| **ORES** | 69.7% | 0.695 |
| **WiSDe** | **80.8%** | **0.880** |
| **WiSDe + ORES** | **82.1%** | **0.886** |

Everything is computed with XGBoost
Best scores are highlighted in bold

[4] Dataset available at http://bit.ly/wiki_spammers.

smaller, and they edit faster than benign users. Regarding edits on talk pages, we have that the majority of the users are not using talk pages (percentage for both benign users and spammers is 69.7%). However, surprisingly, we have that, among users editing talk pages, the talk page edit ratio is higher for spammers (0.2) than for benign users (0.081), and we observe a group of around 303 spammers trying to gain visibility by making numerous edits on talk pages. Finally, username based features contribute to an increase in accuracy prediction by 2.9% (from 77.9% to 80.8%) and Mean Average Precision by 0.019 (from 0.861 to 0.880).

We compared WiSDe with ORES only, as the tool proposed in [44] is no longer used, and Cluebot NG and STiki are explicitly designed for vandalism and not spam. To compare our system with ORES, we considered the edit damaging score. More specifically, given a user and all her edits, we computed both the average and maximum damaging score provided by ORES and used these as features for classification. Results on 10-fold cross-validation with XGBoost (the best performing classifier) are reported in Table 5, as well. As we can see, ORES performances are poor for the task of spammer detection (69.7% of accuracy and mean average precision of 0.695). However, combining our features with ORES further increases the accuracy to 82.1%.

In reality, spam users are greatly outnumbered by benign users. Thus, similarly to what we did in the previous section, we also created an unbalanced dataset to test our system WiSDe by randomly selecting users at a ratio of 10% spammers and 90% of benign users. Then, we performed 10-fold cross-validation and measured the performance by using the area under the ROC curve (AUROC). To deal with class imbalance, we oversampled the minority class in each training set by using SMOTE [98]. We also considered class weighting, but we found that SMOTE is performing the best. Due to the randomness introduced, we repeated each experiment 10 times and averaged the results. Table 6 reports the results for this experiment. As we can see, even with class imbalance, WiSDe reaches a good AUROC of 0.842 (in comparison we have an AUROC of 0.891 for the balanced setting) and significantly improve over ORES (AUROC of 0.736). However, adding ORES features to ours helps to increase the AUROC to 0.864.

**Table 6** WiSDe vs. ORES performance in the unbalanced setting

| | AUROC |
|---|---|
| **ORES** | 0.736 |
| **WiSDe** | **0.842** |
| **WiSDe + ORES** | **0.864** |

Everything is computed by using XGBoost
Best scores are highlighted in bold

### *4.3 Content Quality Protection and User Retention*

As we have seen in this chapter, a lot of research has been done with the aim of maintaining the trustworthiness, legitimacy, and integrity of Wikipedia content. However, one big drawback of deploying anti-vandalism and anti-spam tools is that veteran editors started to suspiciously look at newcomers as potential vandals and rapidly and unexpectedly deleted contributions even from good-faith editors. Many newcomers, in fact, face social barriers [99] preventing them from the integration in the editor community, with the consequence of stop editing after a certain period of time [100, 101]. As Halfaker et al. [102] have pointed out, Wikipedia is not anymore the encyclopedia that anyone can edit but rather *"the encyclopedia that anyone who understands the norms, socializes himself or herself, dodges the impersonal wall of semi-automated rejection, and still wants to voluntarily contribute his or her time and energy can edit."*

The loss of active contributors from any user-generated content community may affect the quantity and quality of content provision not only on the specific community but also on the Web in general. Most importantly, UGC communities mainly survive thanks to the continued participation of their active users who contribute with their content production.

Thus, being able to early predict whether or not a user will become inactive is very valuable for Wikipedia and any other user-generated content community to perform engaging actions on time to keep these users contributing longer. In our related work [90, 91], we addressed the problem of early predict whether or not a Wikipedia editor will become inactive and stop contributing and proposed a predictive model based on users' editing behavior that achieves an AUROC of 0.98 and a precision of 0.99 in predicting inactive users. By comparing the editing behavior of active vs. inactive users, we discovered that active users are more involved in edit wars and positively accept critiques, and edit much more different categories of pages. On the other hand, inactive users have more edits reverted and edit more meta-pages (and in particular *User* pages).

Regarding specific actions for engaging editors, the Wikipedia community considers several steps that can be taken to increase the retention rate such as (1) survey newly registered users to capture user's interests and use them for making relevant editing recommendations and (2) connect editors with similar interests to form meaningful contribution teams. They also developed and deployed a tool, called Snuggle [103], to support newcomers socialization.

## 5 Conclusions

In this chapter, we discussed several types of misinformation that these days exist on the Web such as vandalism, spam, fraudulent reviews, fake news, etc. and provided a survey on how to detect them.

Then, we focused on the specific case study of protecting Wikipedia from misinformation and presented our research on detecting pages to protect and identifying spam users. Our experimental results show that we are able to classify (1) article pages to protect with an accuracy of 92% across multiple languages and (2) spammers from benign users with 80.8% of accuracy and 0.88 mean average precision. Both the methods proposed do not look at edit content and, as a consequence, they are generally applicable to all versions of Wikipedia, not only the English one.

Finally, we discussed a possible solution for newcomers retention, given that many new users do not keep contributing to the encyclopedia because of the tools deployed to fight vandalism.

## References

1. L. Wu, F. Morstatter, X. Hu, H. Liu, Mining misinformation in social media, in *Big Data in Complex and Social Networks* (2016), pp. 123–152
2. K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: a data mining perspective. ACM SIGKDD Explor. Newslett. **19**(1), 22–36 (2017)
3. A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, R. Procter, Detection and resolution of rumours in social media: a survey. ACM Comput. Surv. (CSUR) **51**(2), 32 (2018)
4. S. Kumar, N. Shah, False information on web and social media: a survey (2018). arXiv preprint:1804.08559
5. Vandalism in Wikipedia. http://en.wikipedia.org/wiki/Wikipedia:Vandalism
6. Spam in Wikipedia. http://en.wikipedia.org/wiki/Wikipedia:Spam
7. Cluebot_NG. http://bit.ly/ClueBotNG
8. STiki. http://bit.ly/STiki_tool
9. A. Shrestha, F. Spezzano, M.S. Pera, Who is really affected by fraudulent reviews? an analysis of shilling attacks on recommender systems in real-world scenarios, in *Late-Breaking Results track part of the Twelfth ACM Conference on Recommender Systems (RecSys'18)* (2018)
10. Pew Research Center. http://www.journalism.org/2016/12/15/many-americans-believe-fake-news-is-sowing-confusion/
11. S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online. *Science*, **359**(6380), 1146–1151 (2018)
12. https://www.factcheck.org/2016/11/how-to-spot-fake-news/
13. H. Allcott, M. Gentzkow, Social media and fake news in the 2016 election. J. Econ. Perspect. **31**(2), 211–36 (2017)
14. P. America, *Faking News: Fraudulent News and the Fight for Truth* (2018). https://pen.org/faking-news/
15. A. Zubiaga, E. Kochkina, M. Liakata, R. Procter, M. Lukasik, Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations, in *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11–16, 2016, Osaka, Japan* (2016), pp. 2438–2448
16. S. Kwon, M. Cha, K. Jung, W. Chen, Y. Wang, Aspects of rumor spreading on a microblog network, in *Proceedings of Social Informatics—5th International Conference, SocInfo 2013, Kyoto, Japan, November 25–27, 2013* (2013), pp. 299–308
17. A. Friggeri, L.A. Adamic, D. Eckles, J. Cheng, Rumor cascades, in *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1–4, 2014* (2014)

18. S. Kumar, R. West, J. Leskovec, Disinformation on the web: impact, characteristics, and detection of wikipedia hoaxes, in *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11–15, 2016* (2016), pp. 591–602
19. P. Bajaj, M. Kavidayal, P. Srivastava, M.N. Akhtar, P. Kumaraguru, Disinformation in multimedia annotation: misleading metadata detection on youtube, in *Proceedings of the 2016 ACM Workshop on Vision and Language Integration Meets Multimedia Fusion, iVandL-MM@MM 2016, Amsterdam, Netherlands, October 16, 2016* (2016), pp. 53–61
20. A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, J. Yosinski, Plug and play generative networks: conditional iterative generation of images in latent space, in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017* (2017), pp. 3510–3520
21. E. Gibney, The scientist who spots fake videos. *Nature* (2017)
22. M. Potthast, B. Stein, R. Gerling, Automatic vandalism detection in wikipedia, in *Proceedings of Advances in Information Retrieval, 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008* (2008), pp. 663–668
23. M. Potthast, B. Stein, T. Holfeld, Overview of the 1st international competition on wikipedia vandalism detection, in *CLEF 2010 LABs and Workshops, Notebook Papers, 22–23 September 2010, Padua, Italy* (2010)
24. B.T. Adler, L. de Alfaro, S.M. Mola-Velasco, P. Rosso, A.G. West, Wikipedia vandalism detection: combining natural language, metadata, and reputation features, in *International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)* (2011), pp. 277–288
25. A.G. West, S. Kannan, I. Lee, Detecting wikipedia vandalism via spatio-temporal analysis of revision metadata? in *Proceedings of the Third European Workshop on System Security, EUROSEC 2010, Paris, France, April 13, 2010* (2010), pp. 22–28
26. B.T. Adler, L. de Alfaro, I. Pye, Detecting wikipedia vandalism using wikitrust—lab report for PAN at CLEF 2010, in *CLEF 2010 LABs and Workshops, Notebook Papers, 22–23 September 2010, Padua, Italy* (2010)
27. J. Kiesel, M. Potthast, M. Hagen, B. Stein, Spatio-temporal analysis of reverted wikipedia edits, in *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15–18, 2017* (2017), pp. 122–131
28. S. Heindorf, M. Potthast, B. Stein, G. Engels, Vandalism detection in wikidata, in *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24–28, 2016* (2016), pp. 327–336
29. P. Neis, M. Goetz, A. Zipf, Towards automatic vandalism detection in openstreetmap. ISPRS Int. J. Geo Inf. **1**(3), 315–332 (2012)
30. S. Kumar, F. Spezzano, V.S. Subrahmanian, VEWS: a wikipedia vandal early warning system, in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10–13, 2015* (2015), pp. 607–616
31. G. Stringhini, C. Kruegel, G. Vigna, Detecting spammers on social networks, in *Proceedings of the 26th Annual Computer Security Applications Conference* (2010), pp. 1–9
32. K. Lee, J. Caverlee, S. Webb, Uncovering social spammers: social honeypots + machine learning, in *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2010), pp. 435–442
33. J. Song, S. Lee, J. Kim, Spam filtering in twitter using sender-receiver relationship, in *International Workshop on Recent Advances in Intrusion Detection* (2011), pp. 301–317
34. C. Yang, R.C. Harkreader, G. Gu, Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers, in *International Workshop on Recent Advances in Intrusion Detection*, pp. 318–337 (2011)
35. C. Grier, K. Thomas, V. Paxson, M. Zhang, @ spam: the underground on 140 characters or less, in *Proceedings of the 17th ACM Conference on Computer and Communications Security (CCS)* (2010), pp. 27–37

36. L. Wu, X. Hu, F. Morstatter, H. Liu, Detecting camouflaged content polluters, in *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15–18, 2017* (2017), pp. 696–699
37. X. Hu, J. Tang, H. Gao, H. Liu, Social spammer detection with sentiment information, in *2014 IEEE International Conference on Data Mining (ICDM)* (2014), pp. 180–189
38. S. Lee, J. Kim, Warningbird: detecting suspicious urls in twitter stream. in *19th Annual Network and Distributed System Security Symposium, NDSS 2012, San Diego, California, USA, February 5–8, 2012* (2012)
39. J. Ma, L.K. Saul, S. Savage, G.M. Voelker, Beyond blacklists: learning to detect malicious web sites from suspicious urls, in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28–July 1, 2009* (2009), pp. 1245–1254
40. D.K. McGrath, M. Gupta, Behind phishing: an examination of phisher modi operandi, in *Proceedings of First USENIX Workshop on Large-Scale Exploits and Emergent Threats, LEET '08, San Francisco, CA, USA, April 15, 2008* (2008)
41. C. Cao, J. Caverlee, Detecting spam urls in social media via behavioral analysis, in *Proceedings of Advances in Information Retrieval—37th European Conference on IR Research, ECIR 2015, Vienna, Austria March 29–April 2, 2015* (2015), pp. 703–714
42. D. Antoniades, I. Polakis, G. Kontaxis, E. Athanasopoulos, S. Ioannidis, E.P. Markatos, T. Karagiannis, we.b: the web of short URLs, in *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28–April 1, 2011* (2011), pp. 715–724
43. A.G. West, J. Chang, K. Venkatasubramanian, O. Sokolsky, I. Lee, Link spamming wikipedia for profit, in *CEAS* (2011), pp. 152–161
44. A.G. West, A. Agrawal, P. Baker, B. Exline, I. Lee, Autonomous link spam detection in purely collaborative environments, in *WikiSym* (2011), pp. 91–100
45. N. Jindal, B. Liu, Opinion spam and analysis, in *Proceedings of the International Conference on Web Search and Web Data Mining, WSDM 2008, Palo Alto, California, USA, February 11–12, 2008* (2008), pp. 219–230
46. M. Ott, Y. Choi, C. Cardie, J.T. Hancock, 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, 19–24 June, 2011, Portland, Oregon, USA* (2011), pp. 309–319
47. A. Mukherjee, V. Venkataraman, B. Liu, N.S. Glance, What yelp fake review filter might be doing? in *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA, July 8–11, 2013* (2013)
48. G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, R. Ghosh, Exploiting burstiness in reviews for review spammer detection, in *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA, July 8–11, 2013* (2013)
49. E. Lim, V.A. Nguyen, N. Jindal, B. Liu, H.W. Lauw, Detecting product review spammers using rating behaviors, in *Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26–30, 2010* (2010), pp. 939–948
50. A. Mukherjee, A. Kumar, B. Liu, J. Wang, M. Hsu, M. Castellanos, R. Ghosh, Spotting opinion spammers using behavioral footprints, in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11–14, 2013* (2013), pp. 632–640
51. K.C. Santosh, A. Mukherjee, On the temporal dynamics of opinion spamming: case studies on yelp, in *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11–15, 2016* (2016), pp. 369–379

52. S. Rayana, L. Akoglu, Collective opinion spam detection: Bridging review networks and metadata, in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10–13, 2015* (2015), pp. 985–994
53. G. Wang, S. Xie, B. Liu, P.S. Yu, Review graph based online store review spammer detection, in *Proceedings of the 11th IEEE International Conference on Data Mining, ICDM 2011, Vancouver, BC, Canada, December 11–14, 2011* (2011), pp. 1242–1247
54. B. Hooi, N. Shah, A. Beutel, S. Günnemann, L. Akoglu, M. Kumar, D. Makhija, C. Faloutsos, BIRDNEST: bayesian inference for ratings-fraud detection, in *Proceedings of the 2016 SIAM International Conference on Data Mining, Miami, Florida, USA, May 5–7, 2016* (2016), pp. 495–503
55. S. Kumar, B. Hooi, D. Makhija, M. Kumar, C. Faloutsos, V.S. Subrahmanian, REV2: fraudulent user prediction in rating platforms, in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5–9, 2018* (2018), pp. 333–341
56. A. Beutel, W. Xu, V. Guruswami, C. Palow, C. Faloutsos, Copycatch: stopping group attacks by spotting lockstep behavior in social networks, in *Proceedings of the 22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13–17, 2013* (2013), pp. 19–130
57. S. Hosseinimotlagh, E.E. Papalexakis, Unsupervised content-based identification of fake news articles with tensor decomposition ensembles, in *MIS2: Misinformation and Misbehavior Mining on the Web Workshop held in conjunction with WSDM 2018 Feb 9, 2018—Los Angeles, California, USA, 2018* (2018)
58. M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, B. Stein, A stylometric inquiry into hyperpartisan and fake news. CoRR, abs/1702.05638 (2017)
59. B.D. Horne, S. Adali, This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news (2017). arXiv preprint:1703.09398
60. V. Pérez-Rosas, B. Kleinberg, A. Lefevre, R. Mihalcea, Automatic detection of fake news, in *Proceedings of the 27th International Conference on Computational Linguistics* (2018), pp. 3391–3401
61. Z. Jin, J. Cao, Y. Zhang, J. Zhou, Q. Tian, Novel visual and statistical image features for microblogs news verification. IEEE Trans. Multimedia **19**(3), 598–608 (2017)
62. J. Kim, B. Tabibian, A. Oh, B. Schölkopf, M. Gomez-Rodriguez, Leveraging the crowd to detect and reduce the spread of fake news and misinformation, in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5–9, 2018* (2018), pp. 324–332
63. Z. Jin, J. Cao, Y. Zhang, J. Luo, News verification by exploiting conflicting social viewpoints in microblogs, in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12–17, 2016, Phoenix, Arizona, USA* (2016), pp. 2972–2978
64. L. Wu, H. Liu, Tracing fake-news footprints: characterizing social media messages by how they propagate, in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5–9, 2018* (2018), pp. 637–645
65. N. Ruchansky, S. Seo, Y. Liu, CSI: a hybrid deep model for fake news detection. in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06–10, 2017* (2017), pp. 797–806
66. N. Knauf, J. Fairbanks, N. Fitch, E. Briscoe, Credibility assessment in the news: do we need to read? in *MIS2: Misinformation and Misbehavior Mining on the Web Workshop held in conjunction with WSDM 2018 Feb 9, 2018—Los Angeles, California, USA, 2018* (2018)
67. K. Shu, S. Wang, H. Liu, Beyond news contents: the role of social context for fake news detection, in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11–15, 2019* (2019), pp. 312–320

68. A. Chakraborty, B. Paranjape, S. Kakarla, N. Ganguly, Stop clickbait: detecting and preventing clickbaits in online news media, in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016, San Francisco, CA, USA, August 18–21, 2016* (2016), pp. 9–16
69. M. Potthast, S. Köpsel, B. Stein, M. Hagen, Clickbait detection, in *Proceedings of Advances in Information Retrieval—38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016* (2016), pp. 810–817
70. A. Anand, T. Chakraborty, N. Park, We used neural networks to detect clickbaits: you won't believe what happened next!, in *Proceedings of Advances in Information Retrieval—39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8–13, 2017* (2017) pp. 541–547
71. Y. Chen, N.J. Conroy, V.L. Rubin, Misleading online content: recognizing clickbait as "false news", in *Proceedings of the 2015 ACM Workshop on Multimodal Deception Detection, WMDD@ICMI 2015, Seattle, Washington, USA, November 13, 2015* (2015), pp. 15–19
72. S. Hamidian, M. Diab, Rumor detection and classification for twitter data, in *Proceedings of the Fifth International Conference on Social Media Technologies, Communication, and Informatics (SOTICS)* (2015), pp. 71–77
73. S. Hamidian, M. Diab, Rumor identification and belief investigation on twitter, in *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (2016), pp. 3–8
74. V. Qazvinian, E. Rosengren, D.R. Radev, Q. Mei, Rumor has it: identifying misinformation in microblogs, in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27–31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL* (2011), pp. 1589–1599
75. A. Zubiaga, M. Liakata, R. Procter, Exploiting context for rumour detection in social media, in *International Conference on Social Informatics* (Springer, Berlin, 2017), pp. 109–123
76. L. Wu, J. Li, X. Hu, H. Liu, Gleaning wisdom from the past: early detection of emerging rumors in social media, in *Proceedings of the 2017 SIAM International Conference on Data Mining, Houston, Texas, USA, April 27–29, 2017* (2017), pp. 99–107
77. D. Shah, T. Zaman, Rumors in a network: who's the culprit? IEEE Trans. Infor. Theory **57**(8), 5163–5181 (2011)
78. W. Luo, W.-P. Tay, Finding an infection source under the SIS model, in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26–31, 2013* (2013), pp. 2930–2934
79. W. Dong, W. Zhang, C.W. Tan, Rooting out the rumor culprit from suspects, in *Proceedings of the 2013 IEEE International Symposium on Information Theory, Istanbul, Turkey, July 7–12, 2013* (2013), pp. 2671–2675
80. C. Kang, S. Kraus, C. Molinaro, F. Spezzano, V.S. Subrahmanian, Diffusion centrality: a paradigm to maximize spread in social networks. Artif. Intell. **239**, 70–96 (2016)
81. E. Tacchini, G. Ballarin, M.L. Della Vedova, S. Moret, L. de Alfaro, Some like it hoax: automated fake news detection in social networks. CoRR, abs/1704.07506 (2017)
82. ORES. http://bit.ly/wikipedia_ores
83. ORES API. http://ores.wikimedia.org
84. B.M. Hill, A.D. Shaw, Page protection: another missing dimension of wikipedia research, in *Proceedings of the 11th International Symposium on Open Collaboration, San Francisco, CA, USA, August 19–21, 2015* (2015), pp. 15:1–15:4
85. Pywikibot. https://www.mediawiki.org/wiki/Manual:Pywikibot/protect.py
86. Hoaxes on Wikipedia. https://en.wikipedia.org/wiki/Wikipedia:List_of_hoaxes_on_Wikipedia
87. K. Suyehira, F. Spezzano, Depp: a system for detecting pages to protect in wikipedia, in *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24–28, 2016* (2016), pp. 2081–2084

88. F. Spezzano, K. Suyehira, L.A. Gundala, Detecting pages to protect in wikipedia across multiple languages. Soc. Netw. Anal. Min. **9**(1), 10 (2018)
89. T. Green, F. Spezzano, Spam users identification in wikipedia via editing behavior, in *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15–18, 2017* (2017), pp. 532–535
90. H. Arelli, F. Spezzano, Who will stop contributing?: predicting inactive editors in wikipedia, in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, Sydney, Australia, July 31–August 03, 2017* (2017), pp. 355–358
91. H. Arelli, F. Spezzano, A. Shrestha, Editing behavior analysis for predicting active and inactive users in wikipedia, in *Influence and Behavior Analysis in Social Networks and Social Media* (2019), pp. 127–147
92. Edit War in Wikipedia. http://en.wikipedia.org/wiki/Wikipedia:Editwarring
93. T. Fawcett, An introduction to ROC analysis. Pattern Recogn. Lett. **27**(8), 861–874 (2006)
94. R. Zafarani, H. Liu, 10 bits of surprise: detecting malicious users with minimum information, in *CIKM* (2015), pp. 423–431
95. http://en.wikipedia.org/wiki/Category:Wikipedians_who_are_indefinitely_blocked_for_spamming
96. http://en.wikipedia.org/wiki/Category:Wikipedians_who_are_indefinitely_blocked_for_link-spamming
97. http://en.wikipedia.org/wiki/Special:BlockList
98. N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
99. I. Steinmacher, T. Conte, M.A. Gerosa, D.F. Redmiles, Social barriers faced by newcomers placing their first contribution in open source software projects, in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW 2015, Vancouver, BC, Canada, March 14–18, 2015* (2015), pp. 1379–1392
100. L. Jian, J.K. MacKie-Mason, Why leave wikipedia? in *iConference* (2008)
101. S. Asadi, S. Ghafghazi, H.R. Jamali, Motivating and discouraging factors for wikipedians: the case study of persian wikipedia. Libr. Rev. **62**(4/5), 237–252 (2013)
102. A. Halfaker, R.S. Geiger, J.T. Morgan, J. Riedl, The rise and decline of an open collaboration system: how wikipedia's reaction to popularity is causing its decline. Am. Behav. Sci. **57**(5), 664–688 (2013)
103. A. Halfaker, R.S. Geiger, L.G. Terveen, Snuggle: designing for efficient socialization and ideological critique, in *CHI Conference on Human Factors in Computing Systems, CHI'14, Toronto, ON, Canada—April 26–May 01, 2014* (2014), pp. 311–320