MANIFESTO: a huMAN-centric explainable approach for Fake news spreaders deTectiOn

Orestis Lampridis¹ · Dimitra Karanatsiou¹ · Athena Vakali¹

Received: 6 March 2021 / Accepted: 2 September 2021 / Published online: 11 January 2022

© The Author(s), under exclusive licence to Springer-Verlag GmbH Austria, part of Springer Nature 2021

Abstract

Fake news spreading is strongly connected with the human involvement as individuals tend to fall, adopt and circulate misinformation stories. Until recently, the role of human characteristics in fake news diffusion, in order to deeply understand and fight misinformation patterns, has not been explored to the full extent. This paper suggests a human-centric approach on detecting fake news spreading behavior by building an explainable fake-news-spreader classifier based on psychological and behavioral cues of individuals. Our model achieves promising classification results while offering explanations of human motives and features behind fake news spreading behavior. Moreover, to the best of our knowledge, this is the first study that aims at providing a fully explainable setup that evaluates fake news spreading based on users credibility applied to public discussions aiming to a comprehensive way to combat fake news through human involvement.

Keywords Fake news · Social networks · User profiling · Explainable machine learning

1 Introduction

With the sharp rise of social networks, users are able to access, share, and discuss on almost any kind of online circulated content. The ease and comfort of online social networks (OSNs), encourages individuals to turn on OSNs platforms as the major point for news seeking, updating, discussing, etc [15]. However, lurking in the shadows of

✉ Dimitra Karanatsiou
dimitrakd@csd.auth.gr

Orestis Lampridis
lorestis@csd.auth.gr

Athena Vakali
avakali@csd.auth.gr

¹ Informatics Department, Aristotle University of Thessaloniki, Thessaloniki, Greece

OSNs lies a vast amount of deception, misinformation and disinformation cases. These cases have established the familiar term *fake news* [27]. Unfortunately, OSNs stand out as main sources at which misinformation is spread and platforms, like the popular Twitter, have become the most used fora at which fake news stories exist and thrive. Human ethics and democratic views are at risk since even presidential elections are manipulated [7], and critical health decisions (e.g. covid pandemic fake stories) are blended with dark conspiracy theories¹ [22]. Data science and AI research should dynamically act on these phenomena with robust methods and techniques which will detect and block OSNs misinformation spreading.

The vast majority of the literature has so far detected misinformation at a post-level exploration, tagging of a specific piece of information as a fake news content [2,20,36,39,44,53,55]. On the contrary, users tendency to circulate misinformation has only recently gained attention. Recent work has explored the role of people in misinformation spreading highlighting the role of users' profiles. Other studies have focused on user profile characteristics such as gender, personality, users' psycholinguistic features or, in other words, language use that reflects psychological aspects, as well as sentiment and emotions which are expressed in a user's text [17,48,50] correlated with the probability of being a fake news spreader. Therefore, the human role is doubly crucial on fake news challenge: both in terms of acting as a fake news spreader and being exposed to fake news content.

After a piece of information is detected as fake news, specific anti-hoax platforms(e.g. PolitiFact²) or even major news outlets report that widespread false information has been validated as such in order to inform the public. However, this is not the case with smaller-scale conspiracy theories that could harm the common good or create confusion on the society. Also, users are not informed about their level of vulnerability to misinformation, in terms of the dissemination that is done intentionally or unintentionally in discussions with other users. Characterizing a suspicious piece of text as fake news cannot stand alone effectively if there is no mechanism that could help the human factor to be aware in case a piece of received information is likely to include fake news or intervene to stop their dissemination. Explainable machine learning (ML) consists a proper and state-of-the-art method that ties in with fake news detection [26,35,57]. Indeed, previous work incorporate explainability in the process of understanding why a news post is labeled as fake. For instance, dFEND [49] utilizes the comments made by the users in the fake news post in order to offer the final explanation while GCAN [30] merely uses words found in the source text. Finally, a review in the topic [42] utilizes SHAP [31] in order to find the most important features in fake news detection at post level. However, no previous work made a comprehensive effort to provide a human understandable mechanism that informs users about their exposure to fake news spreading.

A deeper study of users roles and profiles is strongly required in this topic since behavioral human-centric variations largely impact the overall fake news lifecycle (origination, spreading, and virality). The present work is motivated by these fake news

¹ <https://www.forbes.com/sites/brucelee/2020/07/11/face-masks-with-5g-antennas-the-latest-covid-19-coronavirus-conspiracy-theory/#16339ef416f1>.

² <https://www.politifact.com/>.

spreading challenges, with a goal to propose a human-centric and explainable approach for detecting the user profiles which are suspicious for misinformation spreading. Our approach focuses on the authors in order to (a) detect user profiles who are suspicious of misinformation spreading based on their profiles and (b) provide a human understandable mechanism that would inform users about their or others' tendency to fake news spreading. Specifically, our contributions are as next: (a) *build an explainable fake-news-spreader classifier based on psychological and behavioral cues*. We use psychological and behavioral features to train a predictive model for fake news spreaders detection that compared to the state of the art approaches using similar features achieves 4.8–8.14% better performance. Moreover, compared to the currently best performing classifier that uses only language features we achieve comparable performance, indicating that our model has promising results on this task. In addition, to the best of our knowledge, the present work proposes the first explainable ML framework for fake news spreaders detection.

(b) *reveal fake news spreaders features leveraging an explainable ML approach*. By uncovering fake news spreaders human-centric (i.e. social and psychological) characteristics, explanations are offered to expose their implicit behavioral patterns and traits. An initial exploration of social and psychological characteristics for both fake and real news spreaders is followed by an explainable ML approach. Our outcomes indicate that negative sentiment polarity, higher influence or power (clout), lower levels of anxiety, agreeableness and openness as well as specific language use characterize fake news spreading behavior.

(c) *design a novel human-centric framework for detecting suspicious users and misinformation elements on public discussions*. In order to apply the fake news spreader classifier in a real world set-up, we identify users tendency to spread misinformation based on public discussions in a fully explainable and human-comprehensive set up. We focus on two types of users involved in such discussions: users who post a news item or an opinion (seed users) and users who participate in the discussion with comments. To approach the problem at a real setting at which misinformation is employed, Twitter platform is used as a rich source of public opinion and users characteristics [4,10]. Two datasets were generated regarding public discussions related to US 2020 Elections and COVID-19 pandemic, posted from individuals that initiate a live conversation (seed posts). We then use the fake news spreader classifier to label people participating in the conversation of each seed post. In that way, we link each reply that expresses an opinion with the credibility of the author in order to train an interpretable linear model the detects misinformative opinions from suspicious users. This simpler model replicates at 87.61% and 71.00% the fake news spreader classifier respectively at each dataset proposing an explainable set up to help end users understand the model with example-based explanations.

Section 2 provides the background on the basic elements involving fake news principles, the role of human factors in misinformation spreading and the need for explainability and human-centric approaches to combat the devastating fake news phenomena. Section 3 presents the design of our approach on training a fake news spreader classifier, building and annotating a real-life dataset and showcasing and evaluating our explainable model for suspicious users for misinformation spreading detection in public discussions. Results on the experimentation based on our described

approach are highlighted in Sect. 4. Finally, we conclude our paper and discuss future work in Sect. 5.

2 Background and related work

In this section, we provide the background related to the definitions and types of fake news that exist online as well as the most common approaches to detect fake news in OSNs. In addition, we focus on understanding the human factor towards misinformation and the social effects that benefit misinformation. Approaches and models developed for detecting fake news spreaders online are also presented.

2.1 Fake news background

Definitions and types. A general consensus is that the term fake news is a form of news that consists of the deliberate spreading of misinformation either through traditional news outlets or OSNs [3,28]. Definitions regarding this term define that fake news is a news article that is intentionally and verifiably false or misleading by design [16,47]. Fake news can appear in various shapes and forms depending on their characteristics [52]. These range from (i) satire or parody, (ii) misleading content, that include partially truthful information to bring down another individual or entity often used in politics, (iii) false context or selective editing includes paraphrased or out of context information, (iv) impostor content, that comes from a fake source pretending to be credible and push a specific agenda, (v) manipulated content, genuinely deceiving information and (vi) fabricated content, which is novel and completely false content that is fabricated with the intent to deceive its readers and create distress. In this study, we use the term fake news to refer to all the misinformation types mentioned above intentional or unintentional misinformation.

Tackling the fake news challenge. Different approaches are proposed to detect fake news pieces online including:

- *Fact checking:* a manual process in which a piece of information undergoes a process of verification to establish its authenticity and truthfulness [18], with the aid of human (non-) experts or automated software.
- *Credibility checking:* involves the analysis of the headlines of the news post (to examine if there are click-bait characteristics), the trustworthiness of the source where the news post is posted and the reputation of people participating in the discussion for this post [10].
- *Network Structure based identification:* includes the transformation of entities involved i.e users who share a post, and analysis of the relationship between same or different types of nodes to detect fake news spreading behavior [55].
- *Predictive Modeling and ML:* employs different algorithms and features that detect fake news posts achieving sufficient levels of accuracy [34,44,54]. However the main disadvantage is that the models most of the times make successful predictions on posts which belongs to the same topic as the data they were initially trained on (e.g. politics, economy, religion, social issues, celebrity gossip, immigration etc.).

Fake news detection using predictive modeling. As for detection approaches using predictive modeling, two directions related to fake news are identified in existing literature: (a) those characterizing the *post* as fake news, and (b) those characterizing the *user* as fake news spreader. Methodologies aiming at post-level usually involve ML with both supervised and unsupervised models. The best performing models were the random forest and XGB classifiers, statistically tied with high accuracy levels [43]. Unsupervised ML methods like clustering are also considered with lower levels of accuracy [11]. Deep learning models [23,45] also employed achieving state-of-the-art performance with accuracy levels reaching 87–89% [23,45]. Recently, researchers identified the need for explainable models that enhance the process of understanding of why a news post is labeled as fake. For instance, DEFEND [49] utilizes the comments made by the users in the fake news post in order to offer the final explanation while GCAN [30] merely uses words found in the source text. On the other side, user-level fake news spreader detection have not gained so much attention. Recent approaches use a wide variety of features including language, emotion and various other metadata (e.g. followers, followees, tweets) offered in social networks to build predictive models to model fake news spreaders profiles [8,9,17,38,46], as research has shown that the user characteristics play an important role in fake news spreading [17,48,50]. In this paper, we deal with the challenge of fake news spreaders detection based on social and psychological characteristics of users after researching for features importance.

2.2 The role of human factor towards misinformation

As mentioned in the previous section, the social and psychological user characteristics play an important role in fake news adoption and diffusion. Fake news spreading is reinforced by these characteristics, as well as by interactions among users and social phenomena that enable them to hatch.

Human characteristics of fake news spreaders. Zhou and Zafarani [56] make a concise classification between OSN users which include malicious users (i.e. fake news spreaders) and non-malicious users. Malicious users also often include bots. However, not all of these bots are malicious as some of them are made for good purposes such as re-posting real news. Indeed, Vosoughi et al. [51] claim that social bots spread false and real news at a similar rate, which means that fake news spread faster and more broadly because real people have a higher likelihood of sharing them indicating their crucial role in this phenomenon. Human aspects related to cognitive psychology affect the susceptibility and vulnerability of social network users to misinformation [26]. With regards to the psychological aspect of fake news adoption and diffusion, only recently there has been a shift of attention in the literature utilizing the user profiles and psychological patterns of social media users in order to classify them as fake or real news spreaders [6]. Recently, Giachanou et al. [17] showed that personality combined with contextual information have a higher predictive power at classifying fake news spreaders. In addition, [48] found correlations between user profile characteristics and tendency to spread to misinformation and recently employed feature importance aid to understand spreaders characteristics based on these profiles [50].

Credibility, trustworthiness and the vulnerability of users towards misinformation consists of behavioral traits that affect the spread of fake news by humans. *Consistency*, the tendency to believe opinions compatible with previous beliefs, *coherency*, the ability to process whether an information makes sense or not, *credibility*, the trustworthiness of the news source and *acceptability*, the power of the herd that affects an individual to adopt or not a piece of information are characterized as important aspects that affect the adoption and diffusion of fake news by individuals [29]. In fact, it has been shown that users with lower credibility have a higher chance of spreading fake news than more reliable users, as users with low credibility are more vulnerable to adopt and reproduce them [1].

In general, two broad categories of features reflect users' credibility and tendency to spread fake news are utilized by related studies: natural language (NLP) and social setting based features. NLP based features include linguistic vectors as n-grams or tf-idf [46], readability statistics such as number characters, words used etc [8], contextual based vectors such as word2vec and GloVe [17], psycholinguistic cues and syntax based features [46]. With regard to social setting based features, user characteristics including explicit features such as platform related metrics (i.e number of followers, number of followees, and other relevant metrics) and implicit features such as personality, age and gender, are employed [9,50].

This study utilizes social and psychological characteristics of users in order build a fake news spreader detection model that would be able to classify users based on their tendency to spread misinformation.

Social effects that benefit misinformation. As fake news can be shared easily in social networks, the exposure of individuals to misinformation has increased dramatically [35]. The "echo-chamber" effect, and homophily principle in human interactions are social effects and theories that escalate the intensity of this phenomenon. The "echo-chamber" effect [14,37] is the state when people have beliefs or views that are different from those that are common within the broader population, and these people preferably communicate with one another managing to maintain, reinforce, or amplify their beliefs or views. Essentially, people tend to follow other like-minded people while ignoring everyone that has differing beliefs imitating the effect of an echo chamber [40]. In addition, individuals tend to receive information which is recommended by others with similar characteristics in their social networks in interpersonal communication, the tendency to form, the known homophily principle [25,32]. This results in the formations of highly polarized clusters in which social homogeneity is the leading actor for the diffusion of information [12]. In terms of misinformation, it seems that people form connections with others expressing similar bias [21] and consume and articulate same types of information [5]. Moreover, evidence show that spreaders share similar attributes allowing us to distinguish fake news from real news by studying these specific users participated in news dissemination and discussion [55] and, thus, detecting misinformation.

This study is motivated from the above theories applied to public discussions in order to exploit the credibility of others in a user's network and build a model for fake news spreading detection.

Transparency and explainability importance. Early detection of fake news is critical in order to stop their further dissemination. Characterizing a suspicious piece of text

Table 1 Overview of the state of the art methods on fake news spreaders detection

	Features				Explainability
	Natural language			Social setting	
	Linguistic	Readability	Psycholinguistic	User characteristics	
[50]				✓	✓
[17]			✓	✓	
[8]	✓	✓			
[46]	✓				
[9]		✓		✓	
MANIFESTO	✓	✓	✓	✓	✓

as fake news cannot standalone effectively if there is no mechanism that could help humans to understand why the information they read or a discussion they participate in include misinformation in order to stop their further dissemination. Explainable ML is a well established state-of-the-art approach employed in fake news detection [26,35,57]. Previous work incorporates explainable ML techniques in the process of interpreting why a news post is labeled as fake. For instance, DEFEND [49] utilizes the comments made by the users in the fake news post in order to offer the final explanation while GCAN [30] merely uses words found in the source text to explain the final classification. Finally, a review in the topic [42] utilizes SHAP³ in order to find the most important features in fake news classification. However, no previous work proposed an extensional method that is not meant to directly detect or explain important features in fake news detection but rather to help users judge by themselves if a received information contains misinformation.

In this work, we apply advanced explainable ML in order to aid the user in making a more educated final decision with regard to real and false pieces of information focusing on the reputation of users who participate in such discussions.

Table 1 summarizes state of the art work in the fake news spreaders detection challenge with emphasis on features and explainability. Our approach utilizes all feature categories found in the background work except from contextual features that are hard to understand and be explained by explainable ML techniques. Specifically, we employ: (a) **linguistic features**, to utilize language use, transforming the raw text data into a vector representation, (b) **readability features**, to utilize writing style, including number of average words per tweet, number of emojis, number of slang words count, number of (full) capitalized words, number of retweets, number of user mentions, number of hashtags, and number of URLs, (c) **psycholinguistic features**, to utilize opinions and sentiment polarity, employing LIWC sentiment-related scores and emotion vectors, and (d) **user characteristics**, including users personality traits and gender. A detailed explanation of these aspects is given in Sect. 3.1 and Table 2. This study fills the gap of an explainable fake-news-spreader classifier based on psychological and behavioral cues able to be interpreted by proposing a novel human-centric approach.

³ <https://shap.readthedocs.io/>.

Fig. 1 Methodology and application depicting phases A, B and C

3 Detecting fake news and spreaders in Twitter employing explainable ML

In this section, we present the analysis we followed, which is used to feed our algorithms for the explainable fake news spreader detection model. At first, we describe the steps we followed to build a model for fake news spreaders detection. Then, we applied interpretable techniques to reveal fake news spreaders features and understand the patterns of this behavior. After this step, we design a novel human-centric framework for detecting suspicious users and misinformation elements on public discussions, in particular, we create two real-life datasets of public discussions by collecting seed posts and the replies for US elections 2020 and COVID-19 pandemic. We apply our model to annotate the replies of the discussion with the reputation of the author. We train a simple linear model able to replicate the fake news spreader model used able to detect misinformative content by suspicious seed users by using information of user network. We also present a fully explainable approach aiming in human understanding of fake news spreading behavior.

Overall methodology and application. The methodology and application of MANIFESTO has three distinct phases: *phase A*, training of Fake News Spreader classifier and providing explanations, *phase B*, creation of tweet post-replies dataset and annotation of authors profiles with the model of phase A, and *phase C*, training a model for misinformative content by suspicious users. For brevity sake, they will be simply called phase A, B, and C.

3.1 Training a fake news spreader classifier

Phase A—Training of a Fake News Spreader classifier includes the following steps: (a) *Utilizing a ground truth dataset for fake news spreaders.* We utilized “Profiling Fake News Spreaders on Twitter dataset” [41] provided by the pan-clef challenge regarding author profiling. The dataset contains the timelines of users sharing fake news as per PolitiFact and Snopes of 300 users on Twitter, equally divided and labelled as real and fake news spreaders. The task is to determine given a Twitter feed if the user is suspicious to spread fake news and misinformation.

(b) Dataset preprocessing. We filtered out links, usernames, punctuations, Twitter special characters, replaced contractions with their full word counterparts, we lower-cased words and removed stopwords, as basic preprocessing steps.

(c) Features extraction. As already explored in the related work different features categories used to train fake news spreading classifiers. We focus on features related to authors profile reflecting social and psychological aspects of the users. The features used to feed our model were scaled in $[0,1]$ and are summarized in Table 2.

We classify the features into two categories: tabular and textual. Tabular features include readability, sentiment, psycholinguistic, personality and gender features and text include linguistic features. The total number of features in the two categories were initially 1028, (1000 best textual features emerged from a univariate feature selection and 28 tabular features). Since the number of the data samples is lower than the number of features and this may lead to generalization issues, we employed the Recursive Feature Elimination (RFE) strategy to select the best features from textual category, starting with all the features and removing them until there are no substantial changes in accuracy score to decide the optimal number of features. As we can see in Fig. 2, after utilizing the 100 most important features the model’s performance does not improve while with a lower number of features the performance is not stable. As a result, we ended up using top 100 textual features.

Since various explanation methods work differently under the hood when given different kinds of data (text and tabular in our case), we had to create two separate models, one which contains only the tabular data (all features minus the linguistic), to draw the explanations from and one that contains all of the data combined to provide meaningful explanations for fake news spreaders.

(d) Models’ training. We experimented with different classical classifiers as well as a neural network (NN) architecture to build the most appropriate model for fake news spreader detection. In order to find the best hyper-parameters we used 80% of the dataset and the rest 20% was used for testing. After best hyper-parameters searching, we performed a cross validation evaluation by using 10 folds on the whole dataset. Best performing algorithms in total were Random Forest (RF) and Gradient Boosting (GB) and we notice that the NN cannot achieve better performance neither with tabular nor with tabular and text features. We also tested the best algorithms using tabular only, text only and tabular and text features.

Table 2 Features used in out experiments along with their explanation

Category	Features	Explanation
Linguistic	Term Frequency–Inverse Document Frequency (TF-ID) vectors	The raw text data into a vector representation, in the form of the TF-IDF schema.
Readability features that give statistical information regarding the text or the way it is read	Avg word count Emoji count Slang words count Capitalized words count Full capitalized words count Retweets count User mentions count Hashtags count Url count	The count of average words per tweet for a user. The total number of emojis in the user's last 300 tweets. The total number of slang words. The total number of capitalized words (e.g. "Yes"). The total number of fully capitalized words (e.g. "YES"). The total number of retweets. The total number of user mentions. The total number of hashtags. The total number of URLs.
Personality features describing the personality of the user based on the language they use	extraversion agreeableness conscientiousness openness neuroticism	Utilizes a pre-trained model that takes as input various text features from the user's last 300 tweets and outputs a score that denotes the user's Big Five personality traits [24].
	avoidance anxiety	Utilizes a pre-trained model that takes as input various text features from the user's last 300 tweets and outputs a score that denotes the user's anxiety and avoidance score [24].
Gender (male or female)	Gender	The gender of the user as predicted with the usage of a pre-trained a naive bayes model achieving an accuracy of 0.71 based on linguistic and readability features
Psycholinguistic features returned by the LIWC lexicon through the language that the users use	Analytic Clout Authentic Tone	Formal, logical, and hierarchical thinking patterns. The relative social status, confidence, or leadership. Portrayal in an authentic or honest way, humility or vulnerability. Positive and negative emotion as given by LIWC.
Emotion and Sentiment features regarding the emotion and sentiment behind the text	anger fear joy sadness	A score that utilizes the emojis in the user's last 300 tweets. An emoji lexicon is used that assigns to each emoji values for the anger, fear, joy and sadness. Then the values are aggregated for each emotion and the final score is calculated.
	negation Vader score Textblob score	The total number of words in negation (e.g. "not", "no"). The sentiment score as given by the python library VADER. The polarity score as given by the python library TextBlob.

3.2 Creation of tweet post-replies datasets and annotation of authors profiles

Phase B describes the creation of two real-life datasets by collecting seed posts and their replies for US elections 2020 and COVID-19 pandemic, in order to study the effectiveness of our fake news detection approach based on the tendency of authors participating in a discussion to be fake news spreaders. For this reason, we apply the model developed in phase A to annotate the replies of the discussion with the reputation of the author in order to showcase our human-centric explainable approach into a real-life setting. *Data collection.* We used the official newly introduced Twitter API 2⁴

⁴ <https://developer.twitter.com/en/docs/twitter-api/early-access>.

Fig. 2 RFE for optimal number of features selection

to gather posts related to US Elections and COVID-19 pandemic with the hashtags #USElections, #Elections2020 and #coronavirus, #covid, #covid19 respectively. Due to the high volume of posts related to the topic and to ensure that the posts that will be included in our analysis gather a sufficient number of comments from others to create a living discussion. Due to this fact and as an evidence of a human handled account, we included posts of users with more than 5000 followers. The second filter applied is the number of replies a seed post received, and after experimentation we decided to include posts received more than 10 replies to be considered as a discussion. As a result, we gathered 1365 tweets that have more than 10 comments and less than 200 for the US elections dataset and 308 tweets that have more than 10 comments and less than 200 in the COVID-19 dataset.

For the users involved in the discussion, including the seed author and the commenters, we collected their 100 most recent tweets and then feed them into the fake news spreader classifier developed in phase A. Then we labeled all users involved in each conversation with their tendency to be a misinformation spreader. Actually, our model, given a tweets feed of a user determines if the user is keen on spreading fake news, in other words it detects their potential credibility. Posts from users with fake news spreader like profiles are more likely to contain misinformation compared to others. As a result, participants' views on a discussion are characterized by the reputation of the author. We present an explainable approach in order to detect seed posts potentially containing misinformation according to the author credibility utilizing information from the author's network.

3.3 Interpretable model training for misinformation spreading by suspicious users

After phase A and phase B and annotation of users participating in the discussion with the tendency to be fake news spreaders, we train an interpretable linear model that

exploits users' reputation and their expressed opinions to detect seed posts that are likely to contain fake news. We evaluate the performance of our model by comparing the linear model with the model of phase A and by showing a human understandable setting of explanations for these predictions.

Utilizing linear models for interpretability Linear models learn a linear function by using the input features, used for both regression and classification problems. In this task we deal with the challenge where a seed Twitter post is characterized as potentially containing misinformation due to the author's credibility. In classification problems, if the linear function assigns a value which is less than zero as the weighted sum of the input features, then the algorithm predicts the negative class and otherwise predicts the positive class. The most widely used linear algorithm is *logistic regression*. The greatest advantage of these models is that the prediction procedure is simple and allows the interpretation of the model. This is the main reason why linear models see such widespread use in other more sensitive fields than ML such as medication, human sciences, brain research, and numerous other quantitative examination fields [33].

Generally, the approaches to interpret ML models can be put in two categories. First, model-agnostic or model-specific, which depends on whether the internal of the model under inspection is required. Second, local vs global, depending on whether the interpretation regards an individual prediction or is given for the entire system. Our approach falls into the local and model-agnostic category.

To get a local level interpretation from a logistic regression model, we need to multiply the vector fed as input with the weight vector learnt by the linear model. This way we get the weights of the features contained in the instance we want to explain (as features that don't exist in the instance we want to explain would have a value of zero in the input vector).

Training process We train a linear model using as input features the texts of all the replies on each seed post and as ground truth the labels assigned by the fake news spreaders model. We have converted the original post and the replies to a vector representation. In particular, we tested two different representations, the bag of words (BoW) model and the tf-idf vector representation. After the linear model is trained we use it to predict the value of the seed post. The final label given by the linear model is compared with the label assigned by the fake news spreader classifier and we evaluate our model with fidelity measure and a comprehensible explainable setup.

Providing classification explanations In order to provide the explainable set up, we draw influence from the global surrogate method which works by training a simpler interpretable model (e.g. a linear model) in order to approximate the predictions made by a more complex model (e.g. a random forest). In this way, we are able to build a simpler model that replicates the more complex offering interpretation and the ability to draw conclusions about its underlying logic at the same time. To augment the explainability component of our approach we choose to also offer the two closest replies from each class as example based explanations. To do this, we calculate the cosine similarity between the initial tweet's vector representation and all the replies' vector representation. Afterwards, all replies are ranked on descending order based on that cosine similarity and the two that have the highest similarity from each class are chosen. These replies are offered to the user as example-based explanations. These explanation strategies offer specific occurrences of the dataset to help clarify the logic

Table 3 Performances of fake news spreader classifiers including only tabular features and both text and tabular features

	Accuracy	Precision	Recall	f1-score
RF (tabular)	0.72	0.72	0.73	0.72
GB (tabular+text)	0.73	0.75	0.70	0.72

of ML models. As a rule, example-based explanations work better if the features that the model has been trained on can be further utilized to convey more context. This means that data in the form of text excel at this category while for tabular data it is more difficult to address such information in a concise manner. Example-based explanations help people better understand ML models as well as the data that was used to train the ML model under inspection.

4 Experimentation and results

In this section, we present the experimentation and results drawn as emerged from the presented methodology and application in Sect. 3.

4.1 Fake news spreaders classifier evaluation

As described in Sect. 3.1, we build a model for detecting fake news spreaders in OSNs. Results drawn from Table 3 indicate that the model trained with only tabular features with the one trained both with tabular and textual features have similar performances, with the GB which considers both tabular and textual features slightly higher achieving a precision score 0.75. However, since explainable ML techniques cannot work with these combination of data we need to have two different models: one for providing explanations based on tabular data to understand the fake news spreading behavior and another trained with tabular and text data to be used as our final fake news spreader detection model. Since we focus on revealing psychological and behavioral features, we use the RF model trained only with tabular data, without a significant loss in accuracy as emerged from Table 3.

4.2 Revealing fake news spreaders features leveraging an explainable ML approach

After training the fake news spreader classifier we proceed with explaining the classification process. Explanations for fake news spreaders characteristics are provided by utilizing ELI5⁵ and SHAP⁶, the most known and widely used tools for explainable ML [13]. These explanations offer highly useful insights as they show how the model works as a whole, by highlighting the most important features after taking into account

⁵ <https://eli5.readthedocs.io/>.

⁶ <https://shap.readthedocs.io/>.

Fig. 3 Feature importance

all cases. We present the feature importance for the tabular features highlighting the top 20 out of the 28 tabular features in the case of SHAP and top 18 in the case of ELI5 in Fig. 3. We also accounted for different training sample sizes, however we did not observe any differences in features importance.

Although the feature ranking with the two methods is different, they share many similarities considering that both of the 2 top features are in the other's top 10. In the top 12 features, they also share the same 9 features, albeit in slightly different ranking. By inspecting SHAP's summary plot in Fig. 3b, we observe high values of polarity score and tone that affect the prediction negatively (contributing to the “real news spreader” class) while low values affect the prediction positively (contributing to the “fake news spreader” class). This means that negative sentiment implies someone is a fake news spreader while positive sentiment implies the opposite. Moreover, high Clout contribute to fake news spreading behavior, while users with lower Clout contribute to real news spreading behavior. With regard to the readability features, high levels of capitalized count generally denote fake news spreading behavior while low levels real news spreading. As for personality features, we can deduce that users with high levels of anxiety, agreeableness and openness generally contribute negatively to the prediction, while lower levels of these features contribute positively. No significant impact on either side can be seen by looking at the conscientiousness, neuroticism or avoidance features.

4.3 Designing a novel human-centric framework for detecting suspicious users and misinformation elements on public discussions

We applied the process described in Section 3.3 on two datasets of public discussions in order to identify users suspicious for misinformation spreading based on public discussions in a fully explainable and human-comprehensive set up. We evaluate the

Table 4 Fidelity average for both datasets and both vector representations

Dataset	Fidelity	
	BOW	TF-IDF
US elections	0.85	0.88
COVID-19	0.70	0.71

linear white box both quantitatively by comparing it to the black-box model for fake news spreaders detection described in Section 3.1 and qualitatively by presenting the explanations on representative examples.

Model's evaluation and explanations.

- *Black-box vs white-box model.* We use fidelity to evaluate our linear model approach compared with the classifications made on the initial post by the fake news spreader classifier. Fidelity is the degree in which the simpler model under inspection can be used to precisely approximate the predictions of a more complex model [19,33]. It is an appropriate measure to evaluate if the linear model is able to accurately classify a piece of text as possibly containing misinformation based on the reputation of authors participating in each discussion. To have a more fair calculation for the fidelity, we excluded the initial posts that had the same label in all the replies. We calculate the fidelity for the rest of the sentences, fidelity equals to 1 if the linear model and fake news spreading classifier agree and 0 otherwise. Finally, the average score were computed for all sentences. Results in Table 4 suggest that fidelity achieves 88.00% of agreement among the fake news spreader classifier and the linear model used Tf-Idf vector. This means that the simpler linear model is able to accurately predict the same label with very high success imitating the more complex fake news spreader classifier. Moreover, as for the prediction accuracy, we can see that the linear model has an overall good performance with respect to learning the fake news spreader classifier as the curve for the ROC curve tends to reach close to the top left corner and respectively for the precision recall curve as it tends to reach the top right corner, as seen in Fig. 4.
- *Qualitative evaluation and explanations.* We qualitatively evaluate our approach by presenting the explanations given by our model. We selected 2 representative examples from suspicious users class along with the latent replies from credible and suspicious users as classified by the linear model. We also present the top features along with their weights as assigned by the linear model (including the bias/intercept). In Table 5, the two examples are classified as suspicious to spread to fake news. As for the first one, the author is making appalling remarks on both presidential candidates, while also making personal and subjective attacks. Replies from real news spreaders state that these accusations are not proven and provide real facts. Replies from fake news seem to agree with the author. As for the features we can see words with positive meaning such as “peace”, push the classification towards the real news (since the real news class is 0), while negative words such as “clown” push it to-wards fake news. These examples show that MANIFESTO is able to give a clueless reader insights about a post they read drawn from the discussion from others could potentially contain fake news by offering

Fig. 4 Accuracy curves of the linear model

them the two closest replies from each class as well as the top features to aid in their judgement and help them better understand and evaluate the tendency to fake news consumption. As for the COVID-19 dataset, we present the examples in Table 6. The first example confuses coronavirus with electoral fraud, with reference to misinformation from within. Short answers from trusted users present the logical voice and reassure while from unreliable users opinions related to electoral fraud and other conspiracy theories are reported. Although the tweet itself would not be qualitatively evaluated as a product of misinformation, the model shows that references to the election result tend to push the categorization towards the false news class. The second example throws rebukes at a public figure. Responses from credible users indicate either that these views are terrifying or they are trying to provide supporting arguments. On the contrary, suspicious users agree with reprimanding and following extremist views.

5 Discussion and conclusions

Comparison with the state of the art. We compare our approach with previous work regarding fake news spreader detection. As for the detection model, we achieve 0.73 accuracy score using text-related features that reflect user psychology and behavior, which is comparable with the state-of-the-art performance achieved by Buda and Bolonyai [8]. They use only textual related features without user profile characteristics, however with the rich features we employed we are able to extract meaningful explanations about the implicit human characteristics of fake news spreaders. Other approaches using deep learning methods with feature vectors such as Saeed et al [46] achieved an accuracy of 0.700 proving that our model outperformed by 4.28%. In addition, Cardaioli et al. [9] utilize similar features in their approach (i.e emotion, personality, language and stylistic etc) achieved an accuracy score of 0.675. Our method reached 8.14% better results, indicating that our results are promising and comparable with the state-of-the-art, hence our model accurately can detect fake news spreader like profiles on Twitter with an explainable ML approach. In addition, our approach

Table 5 Explanations of suspicious users posts for the US elections dataset

Seed post	Replies from credible users	Replies from suspicious users	Top features	Weights
Newsmax host warns Democrats will “round up all the Republicans” into “COVID internment camps”. There’s so much talk about the “election fraud” rhetoric leading to the murderous mob, but this is the kind of shit that led up to it and it’s pumped out by conservative media all day every day	How many Dem election losses are blamed on “voter suppression”? A different way of saying election fraud	Bringing Israel and these Arab countries together is a miracle for me	Bias	-0.085
	Alleged you said not proven so why bring him in and trump he is racist, and baffling buffoon... so clown is better	Absolute clown	Bringing	0.112
			Trump	-0.099
			Peace	-0.083
			Clown	0.081
Has anyone in history campaigned more intensely in more places in a short amount of time than @realDonaldTrump? This guy doesn’t need the money nor power. He’s doing it because he loves the USA!	The only thing Trump loves is himself, power, attention, and money. He certainly has no love or respect for America. Patriots are voting him out on Tuesday	He’s doing it for money and power	Bias	0.056
	It’s like a closed fist husband, standing over a woman called Liberty who is huddled begging for her life crying in the corner with two black eyes, trying to protect her children from harm and hungerwhile he tells her he loves her	Thankful and grateful for our president trump	Doing	0.143
			Power	0.029
			Loves	-0.014
			Money	0.011

Table 6 Explanations of suspicious users posts for the COVID-19 dataset

Seed post	Replies from credible users	Replies from suspicious users	Top features	Weights
Newsmax host warns Democrats will "round up all the Republicans" into "COVID interment camps". There's so much talk about the "election fraud" rhetoric leading to the murderous mob, but this is the kind of shit that led up to it and it's pumped out by conservative media all day every day	I feel as though so many Republicans are genuinely worried about this kind of stuff, and there's no way to convince them that their lives are not in danger/ the government isn't taking over everything no matter how hard you try	How many Dem election losses are blamed on "voter suppression"? A different way of saying election fraud	Bias	0.158
	I was thinking about the "government buildings" bit in the warnings. State health departments should get a head's up	Just the Republican QANon Enablers who PROMOTED Election Fraud, Sedition, Treason and Inciting a MINDLESS ZOMBIE Neo-Nazi Supremacist Mob to Insurrection. So if that means ALL REPUBLICANS, So be it...	Shit	0.079
			Election Kind Fraud	0.077 - 0.059 0.049

Table 6 continued

Seed post	Replies from credible users	Replies from suspicious users	Top features	Weights
The worst news of this week is that Biden has appointed Andy Slavitt to his covid task force. Slavitt is a REALLY bad guy. A charlatan, and a habitual liar and self-promoter who has spread terror for his own aggrandizement. Here is selling a mask he claims "deactivates" covid	How do people listen to that stuff its just 24/7 fear porn	Most of his picks have been good so far. No one's going to bat 1.000. Every fucking person Trump appointed was a piece of shit, so don't flip out about one controversial pick from Biden and pull the "both sides are the same" bullshit I see many are doing on here	Bias	-0.098
	It seems that his remit is limited (temporary) and is confined to communications/messaging around vaccines and coordination in getting vaccines out. We will see how effective he will be at this. Giving even bad actors the benefit of the doubt for the moment	Biden will pick a cabinet of the usual suspects and America will suffer	Biden	0.104
			Covid	-0.074
			Appointed	0.051
			Bad	-0.026

is the first comprehensive effort to explain human motives and features behind fake news spreading behavior.

Human-centric set up for misinformation detection. As shown from our results, we showcase that without loss in accuracy our method builds an explainable setup enhancing human understanding of fake news spreading behavior. Specifically, our approach focus on authors who based on their profiles are classified as suspicious to circulate misinformation on public discussions. Following a conversation on Twitter platform, a simpler model is able to replicate the tendency of users to misinformation while presenting the closest replies that are closest to the seed post and belong to both classes, can help the end user to evaluate the source of the information as well as the quality of the discussion that is conducted in relation to whether it includes misinformation. To the best of our knowledge, this is the first study that aims to provide a fully explainable setup that evaluates misinformation spreading based on users credibility. This approach is dataset agnostic, meaning as it can be applied to any topic based only on authors' reputation in a discussion and their opinion, aiming to a comprehensive way to combat fake news through human involvement.

Limitations and future work. Limitations also exist in our approach. Firstly, the performance of MANIFESTO linear model and consequently the quality of explanations relies on the performance of the fake news spreader classifier. In this line, a creation of a novel larger fake news spreader dataset which also contains more rich information regarding other explicit Twitter metadata (e.g. user description, user picture, number of followers, number of favorites etc.) would be of great first step to further improve the fake news spreader classifier. Moreover, one possible direction is to research for other valuable implicit features from Twitter profiles affecting fake news spreading behavior, such as the educational background of a user. As for the explainable set up, as a future direction we also mention the classification of suspicious users to positive, negative or neutrals towards a piece of information, to provide more information to the final user. Finally, a human evaluation of MANIFESTO (e.g. via crowd-sourcing) would be definitively required for assessing the quality of the explanations and improve their quality.

Acknowledgements This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship, and Innovation, under the call RESEARCH - CREATE - INNOVATE (Project Codes: T1EDK-03052, T2EDK-03898).

References

1. Abbasi MA, Liu H (2013) Measuring user credibility in social media. In: Greenberg AM, Kennedy WG, Bos ND (eds) Social computing. Behavioral-cultural modeling and prediction. Springer, Berlin, pp 441–448
2. Ahmed H, Traore I, Saad S (2017) Detection of online fake news using n-gram analysis and machine learning techniques. In: Traore I, Woungang I, Awad A (eds) Intelligent, secure, and dependable systems in distributed and cloud environments. Springer, Berlin, pp 127–138
3. Allcott H, Gentzkow M (2017) Social media and fake news in the 2016 election. J Econ Perspect 31:211–236. <https://doi.org/10.1257/jep.31.2.211>
4. Atodiresei CS, Tănăselea A, Iftene A (2018) Identifying fake news and fake users on twitter. Proc Comput Sci 126:451–461

5. Bessi A, Petroni F, Del Vicario M, Zollo F, Anagnostopoulos A, Scala A, Caldarelli G, Quattrociocchi W (2015) Viral misinformation: the role of homophily and polarization. In: Proceedings of the 24th international conference on World Wide Web, pp 355–356
6. Bevendorff J, Ghanem B, Giachanou A, Kestemont M, Manjavacas E, Markov I, Mayerl M, Potthast M, Rangel F, Rosso P, Specht G, Stamatatos E, Stein B, Wiegmann M, Zangerle E (2020) Overview of pan 2020: authorship verification, celebrity profiling, profiling fake news spreaders on twitter, and style change detection. In: Arampatzis A, Kanoulas E, Tsikrika T, Vrochidis S, Joho H, Lioma C, Eickhoff C, Névél A, Cappellato L, Ferro N (eds) Experimental IR meets multilinguality, multimodality, and interaction, pp 372–383
7. Bovet A, Makse HA (2019) Influence of fake news in twitter during the 2016 us presidential election. *Nat Commun* 10(1):1–14
8. Buda J, Bolonyai F (2020) An ensemble model using n-grams and statistical features to identify fake news spreaders on twitter. In: CLEF
9. Cardaioli M, Ceconello S, Conti M, Pajola L, Turrin F (2020) Fake news spreaders profiling through behavioural analysis. In: CLEF
10. Castillo C, Mendoza M, Poblete B (2011) Information credibility on twitter. In: Proceedings of the 20th international conference on world wide web, association for computing machinery, WWW '11, p 675–684. <https://doi.org/10.1145/1963405.1963500>
11. Conroy NK, Rubin VL, Chen Y (2015) Automatic deception detection: methods for finding fake news. *Proc Assoc Inf Sci Technol* 52(1):1–4. <https://doi.org/10.1002/prat.2015.145052010082>
12. Del Vicario M, Bessi A, Zollo F, Petroni F, Scala A, Caldarelli G, Stanley HE, Quattrociocchi W (2016) The spreading of misinformation online. *Proc Natl Acad Sci* 113(3):554–559. <https://doi.org/10.1073/pnas.1517441113>
13. Došilović FK, Brčić M, Hlupić N (2018) Explainable artificial intelligence: a survey. In: 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO). IEEE, pp. 0210–0215
14. Flaxman S, Goel S, Rao JM (2016) Filter bubbles, echo chambers, and online news consumption. *Public Opin Q* 80:298–320
15. Flintham M, Karner C, Bachour K, Creswick H, Gupta N, Moran S (2018) Falling for fake news: investigating the consumption of news via social media. In: Proceedings of the 2018 CHI conference on human factors in computing systems, pp 1–10
16. Gelfert A (2018) Fake news: a definition. *Informal Logic* 38(1):84–117. <https://doi.org/10.22329/il.v38i1.5068>
17. Giachanou A, Rissola E, Ghanem B, Crestani F, Rosso P (2020) The role of personality and linguistic patterns in discriminating between fake news spreaders and fact checkers, pp 181–192. https://doi.org/10.1007/978-3-030-51310-8_17
18. Graves L, Amazeen M (2019) Fact-checking as idea and practice in journalism. Oxford University Press, Oxford
19. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D (2019) A survey of methods for explaining black box models. *ACM Comput Surv* 51(5):93
20. Guo C, Cao J, Zhang X, Shu K, Yu M (2019) Exploiting emotions for fake news detection on social media
21. Hakobyan Z, Koulouvatianos C (2019) Populism and polarization in social media without fake news: the vicious circle of biases, beliefs and network homophily
22. Huang B, Carley KM (2020) Disinformation and misinformation on twitter during the novel coronavirus outbreak. arXiv preprint [arXiv:2006.04278](https://arxiv.org/abs/2006.04278)
23. Kaliyar RK, Goswami A, Narang P (2021) Deepfake: improving fake news detection using tensor decomposition-based deep neural network. *J Supercomput* 77(2):1015–1037
24. Karanatsiou D, Sermpetis P, Gruda J, Kafetsios K, Dimitriadis I, Vakali A (2020) My tweets bring all the traits to the yard: predicting personality and relational traits in online social networks. arXiv preprint [arXiv:2009.10802](https://arxiv.org/abs/2009.10802)
25. Kossinets G, Watts DJ (2009) Origins of homophily in an evolving social network. *Am J Sociol* 115(2):405–450
26. Kumar KK, Geethakumari G (2014) Detecting misinformation in online social networks using cognitive psychology. *HCIS* 4(1):1–22
27. Lazer DM, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM, Menczer F, Metzger MJ, Nyhan B, Pennycook G, Rothschild D et al (2018a) The science of fake news. *Science* 359(6380):1094–1096

28. Lazer DMJ, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM, Menczer F, Metzger MJ, Nyhan B, Pennycook G, Rothschild D, Schudson M, Sloman SA, Sunstein CR, Thorson EA, Watts DJ, Zittrain JL (2018b) The science of fake news. *Science* 359(6380):1094–1096. <https://doi.org/10.1126/science.aao2998>
29. Lewandowsky S, Ecker UKH, Seifert CM, Schwarz N, Cook J (2012) Misinformation and its correction: continued influence and successful debiasing. *Psychol Sci Public Interest* 13(3):106–131. <https://doi.org/10.1177/1529100612451018>
30. Lu YJ, Li CT (2020) Gcan: graph-aware co-attention networks for explainable fake news detection on social media
31. Lundberg S, Lee SI (2017) A unified approach to interpreting model predictions
32. McCroskey JC, Richmond VP, Daly JA (1975) The development of a measure of perceived homophily in interpersonal communication. *Hum Commun Res* 1(4):323–332
33. Molnar C (2019) Interpretable machine learning. <https://christophm.github.io/interpretable-ml-book/>
34. Oraby S, Reed L, Compton R, Riloff E, Walker M, Whittaker S (2015) And that's a fact: distinguishing factual and emotional argumentation in online dialogue. In: Proceedings of the 2nd workshop on argumentation mining. Association for Computational Linguistics, pp 116–126. <https://doi.org/10.3115/v1/W15-0515>
35. Pennycook G, Cannon TD, Rand DG (2018) Prior exposure increases perceived accuracy of fake news. *J Exp Psychol Gen* 147(12):1865
36. Pérez-Rosas V, Kleinberg B, Lefevre A, Mihalcea R (2018) Automatic detection of fake news. In: Proceedings of the 27th international conference on computational linguistics. Association for Computational Linguistics, pp 3391–3401
37. Petrov A, Proncheva O (2018) Modeling propaganda battle: decision-making, homophily, and echo chambers. In: Conference on artificial intelligence and natural language. Springer, pp 197–209
38. Pizarro J (2020) Profiling bots and fake news spreaders at pan'19 and pan'20: bots and gender profiling 2019, profiling fake news spreaders on twitter 2020. In: 2020 IEEE 7th international conference on data science and advanced analytics (DSAA). IEEE, pp 626–630
39. Potthast M, Kiesel J, Reinartz K, Bevendorff J, Stein B (2018) A stylometric inquiry into hyperpartisan and fake news. In: Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long Papers). Association for Computational Linguistics, pp 231–240. <https://doi.org/10.18653/v1/P18-1022>
40. Quattrocioni W, Scala A, Sunstein C (2019) Echo chambers on facebook
41. Rangel F, Rosso P, Ghanem B, Giachanou A (2020) Profiling fake news spreaders on twitter. <https://doi.org/10.5281/zenodo.4039435>
42. Reis JCS, Correia A, Murai F, Veloso A, Benevenuto F (2019) Explainable machine learning for fake news detection. In: Proceedings of the 10th ACM conference on web science. Association for Computing Machinery, WebSci '19, p 17–26. <https://doi.org/10.1145/3292522.3326027>
43. Reis JCS, Correia A, Murai F, Veloso A, Benevenuto F (2019) Supervised learning for fake news detection. *IEEE Intell Syst* 34(2):76–81
44. Rubin V, Conroy N, Chen Y (2015) Towards news verification: deception detection methods for news discourse. <https://doi.org/10.13140/2.1.4822.8166>
45. Ruchansky N, Seo S, Liu Y (2017) Csi: a hybrid deep model for fake news detection. In: Proceedings of the 2017 ACM conference on information and knowledge management, pp 797–806
46. Saeed U, Fahim H, Shirazi F (2020) Profiling fake news spreaders on twitter. In: CLEF (Working notes)
47. Shu K, Sliva A, Wang S, Tang J, Liu H (2017) Fake news detection on social media: a data mining perspective. *SIGKDD Explor Newsl* 19(1):22–36. <https://doi.org/10.1145/3137597.3137600>
48. Shu K, Wang S, Liu H (2018) Understanding user profiles on social media for fake news detection. In: 2018 IEEE conference on multimedia information processing and retrieval (MIPR), pp 430–435
49. Shu K, Cui L, Wang S, Lee D, Liu H (2019a) Defend: explainable fake news detection. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, association for computing machinery, KDD '19, pp. 395–405. <https://doi.org/10.1145/3292500.3330935>
50. Shu K, Zhou X, Wang S, Zafarani R, Liu H (2019b) The role of user profile for fake news detection. 1904.13355
51. Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. *Science* 359(6380):1146–1151. <https://doi.org/10.1126/science.aap9559>
52. Wardle C (2017) Fake news. It's complicated. <https://firstdraftnews.org/latest/fake-news-complicated/>. Accessed 28 Aug 2020

53. Yang F, Liu Y, Yu X, Yang M (2012) Automatic detection of rumor on sina weibo. Association for Computing Machinery, New York, MDS '12. <https://doi.org/10.1145/2350190.2350203>
54. Zhang H, Fan Z, Zheng J, Liu Q (2012) An improving deception detection method in computer-mediated communication. J Netw. <https://doi.org/10.4304/jnw.7.11.1811-1816>
55. Zhou X, Zafarani R (2019) Network-based fake news detection: a pattern-driven approach. ACM SIGKDD Explor Newsl 21:48–60. <https://doi.org/10.1145/3373464.3373473>
56. Zhou X, Zafarani R (2020) A survey of fake news: fundamental theories, detection methods, and opportunities. ACM Comput Surv. <https://doi.org/10.1145/3395046>
57. Zhou X, Jain A, Phoha VV, Zafarani R (2020) Fake news early detection: a theory-driven model. Digital Threats Res Pract. <https://doi.org/10.1145/3377478>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.