Countering Fake News Technically – Detection and Countermeasure Approaches to Support Users

7

Katrin Hartwig and Christian Reuter

Abstract

The importance of dealing with fake news has increased in both political and social contexts: While existing studies mainly focus on how to detect and label fake news, approaches to help users make their own assessments are largely lacking. This article presents existing black-box and white-box approaches and compares advantages and disadvantages. In particular, white-box approaches show promise in counteracting reactance, while black-box approaches detect fake news with much greater accuracy. We also present the browser plugin TrustyTweet, which we developed to help users evaluate tweets on Twitter by displaying politically neutral and intuitive warnings without generating reactance.

7.1 Introduction

For some time now, social networks such as Facebook and Twitter have increasingly served as important sources of news and information. The result is a dissemination of information that is partially independent of professional journalism. The large amounts of data and information available can be overwhelming. In this context, the term "information overload" was coined (Kaufhold et al. 2020). At the same time, it facilitates the dissemination of dubious or fake content. Steinebach et al. (2020) cite "high speed, reciprocity, low cost, anonymity, mass dissemination, fitfulness, and invisibility" as characteristics that

K. Hartwig (✉) · C. Reuter
Lehrstuhl Wissenschaft und Technik für Frieden und Sicherheit, Technische Universität Darmstadt, Darmstadt, Germany
e-mail: hartwig@peasec.tu-darmstadt.de; reuter@peasec.tu-darmstadt.de

P. Klimczak, T. Zoglauer (eds.), *Truth and Fake in the Post-Factual Digital Age*,
https://doi.org/10.1007/978-3-658-40406-2_7

favor the spread of disinformation on the Internet and especially in social networks. Furthermore, similar phenomena such as the spread of false rumors or clickbaiting can also occur in professional journalism, favoured by the highly attention-based online market.

Since the 2016 presidential election in the United States, the term fake news has become widespread and has been taken up both in academic contexts and in public debates. Fake news are defined by the EU Commission as "all forms of false, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm" (European Commission 2018). Allcott and Gentzkow (2017, p. 213) define fake news as "news articles that are intentionally and verifiably false, and could mislead readers." Studies have shown that fake news often results from minor changes in the wording, so that the basic sentiment changes, for example, rather than being completely made up (Rashkin et al. 2017).

In Germany, too, the 2017 federal elections were accompanied by discussions about the influence of fake news. However, the research results of a study by Sängerlaub (2017) show that there was no significant fake news during the election campaign that would have influenced the election results. These observations suggest that the public's perception of fake news is different from its actual influence. People often find it difficult to distinguish between fake news and true news, as hardly any fake news is completely false and true news can also contain errors (cf. Potthast et al. 2018).

In addition, even more recent events are accompanied by a flood of misinformation. In particular, problematic clips on the spread of the coronavirus have been called out hundreds of thousands of times on the video platform TikTok. To counteract this, TikTok users are "increasingly reminded to report content" (Breithut 2020). Videos with misleading information are deleted by the company accordingly.

Recent research continues to show that only a limited number of individuals are actually vulnerable to being influenced by fake news (Dutton and Fernandez 2019). A Twitter analysis in the US found that "only 1% of users were exposed to 80% of fake news, and 0.1% of users were responsible for sharing 80% of fake news" (Grinberg et al. 2019). Although the actual impact of fake news is still a controversial topic and research suggests that only a few users are susceptible to it, large parts of the population seem to have already encountered fake news. A representative survey in Germany from 2017 shows that fake news plays a significant role in the perception of the population. 48% stated that they had already experienced fake news. Furthermore, 84% were of the opinion that fake news posed a danger and could manipulate the opinion of the population. 23% stated that they had already deleted or reported fake news. In contrast, only 2% said they had ever created fake news themselves (Reuter et al. 2019). An overview of the results is shown in Fig. 7.1.

In summary, fake news can certainly have negative effects, for example on democracy and public trust (Zhou et al. 2019). In fact, there have already been cases where the spread of fake news has caused significant damage. In 2013, for example, a fake tweet from the hacked account of the US news agency Associated Press caused $130 billion in stock market damage, falsely reporting explosions at the White House (Rapoza 2017). Further,

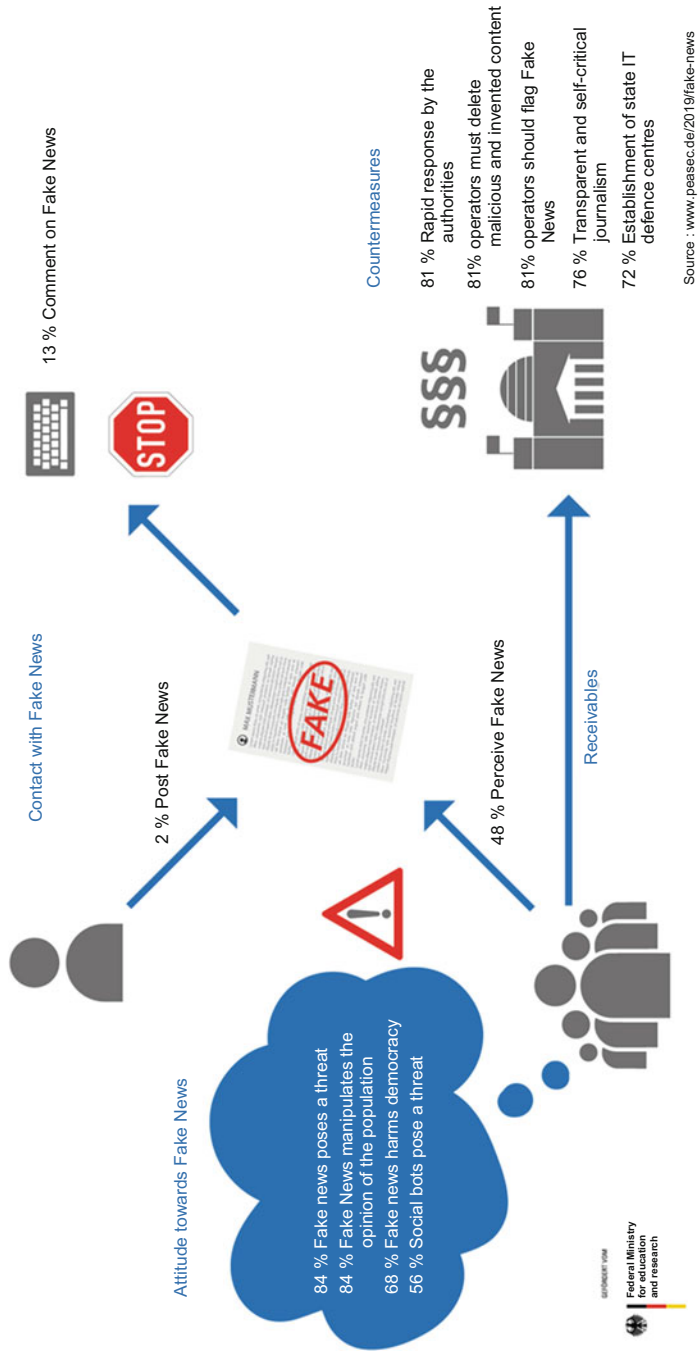

Fig. 7.1 Results of the representative survey on the perception of fake news. (Reuter et al. 2019)

Table 7.1 Steps for technical support in dealing with fake news

Steps	1. Detection	2. Countermeasure approaches
Description	Detecting disinformation; for example, identifying from a set of tweets those tweets that contain fake news	Take measures to protect users from the effects of fake news and empower them to evaluate content themselves

fake news of the #PizzaGate conspiracy theory led to a shooting at a pizzeria in Washington D.C. (Aisch et al. 2016).

Technical solutions for dealing with fake news, especially in social networks, have great potential to counteract the influence of fake news with less user effort. In principle, two steps are necessary in the development of technical support approaches for dealing with fake news: Detect fake news and take countermeasures to protect and support users (Potthast et al. 2018). These are explained in more detail in Table 7.1.

It is also important to consider who is ultimately responsible. The representative study by Reuter et al. (2019) investigated the opinions of the German population on how to deal with fake news. Among other things, participants were asked to rate the following suggestions for dealing with fake news on a five-point Likert scale: quick reactions by the authorities, operators must delete malicious and invented content, operators should flag fake news, transparent and self-critical journalism, and the establishment of state IT defense centers. Most participants said they agreed with all the proposed measures. The idea of setting up state IT defence centres to combat fake news showed the lowest level of acceptance (72%) compared to the other items (Reuter et al. 2019).

7.2 Detection of Fake News in Social Media

7.2.1 Approaches

Since, according to Vosoughi et al. (2018), fake news spread faster than true news, interdisciplinary approaches are essential to address the complex challenges involved. Various methods already exist to detect fake news on social media. For example, platforms can allow their users to report suspicious content. Furthermore, professional fact checkers can manually verify or refute the reported content. In addition, the research field of automated fake news detection is growing through technical solutions, such as style-based fake news detection, propagation-based and context-based fake news detection (Potthast et al. 2018; Zhou et al. 2019).

A good overview of common detection approaches for fake news is provided by Steinebach et al. (2020). The authors distinguish between the detection of misinformation regarding texts, images and bots. Zhang and Ghorbani (2020) further differentiate automatic detection methods according to three categories – *component-based, data mining-based* and *implementation-based* approaches. In this context, component-based detection

methods examine, for example, the authors of fake news or users of social media based on sentiment analysis. Sentiment analysis belongs to the field of text mining and uses signal words, for example, to automatically investigate which sentiments and moods prevail in texts by certain authors. Furthermore, component-based detection methods examine news content on the basis of linguistic (e.g., particularly many exclamation marks), semantic (e.g., particularly attention-grabbing titles that conflict with the body of the text in terms of content), knowledge-based (e.g., websites that use expert knowledge), or style-based (e.g., writing style with a particularly high number of emotional words) features as well as the social context on the basis of user network analyses or distribution patterns. The category of data mining-based detection methods, on the other hand, distinguishes supervised and unsupervised learning. Further, the category of implementation-based approaches distinguishes real-time and offline detection of fake news (cf. Zhang and Ghorbani 2020). The categorization of fake news detection methods can be found in Fig. 7.2.

Many approaches focus on characteristics of text content (Granik and Mesyura 2017; Gravanis et al. 2019; Hanselowski et al. 2019b; Potthast et al. 2018; Rashkin et al. 2017; Zhou et al. 2019). The annotated corpus of Hanselowski et al. (2019a) provides a foundation for machine learning approaches to automated fact checking. Others study user interaction (Long et al. 2017; Ruchansky et al. 2017; Shu et al. 2019b; Tacchini et al. 2017) or content propagation within social networks (Monti et al. 2019; Shu et al. 2019a; Wu and Liu 2018). Other work addresses the relationship of the headline to the body of the text (Bourgonje et al. 2018), argumentation (Sethi 2017), and conflicting perspectives on a topic (Jin et al. 2016).

Following existing approaches for identifying spam messages, Naive Bayes classifiers are often used for the detection and probability calculation of fake news. Here, objects are assigned to a class (e.g., (a) fake news or (b) correct information) that they are most likely to resemble, based on *Bayes'* mathematical *theorem*. Since articles that contain fake news often share the same word groups, Naive Bayes classifiers can be used to calculate the probability that articles contain fake news (Granik and Mesyura 2017). Both Pérez-Rosas et al. (2017) and Potthast et al. (2018) resort to linguistic and semantic features (e.g., certain N-grams, sentence and word proportions) for fake news detection. In this context, Potthast et al. (2018) focus in particular on stylistic features for news containing left- or right-wing extremist content. Here, it is noticeable that despite very different political orientations, the writing styles used are very similar. Furthermore, it becomes apparent that superlatives and exaggerations are increasingly used in fake news (Rashkin et al. 2017).

Algorithmic machine learning approaches, e.g. from Perspective API, are used to recognize certain speech patterns that are essential for detecting fake news. These check statements and messages for short sentences and certain tenses, for example. Fact-checking websites such as PolitiFact rank verified articles on a scale scaled from "true" to "absolutely false" (Rashkin et al. 2017, p. 2931). Another approach by Gravanis et al. (2019) describes that identifying fake news requires a tool that can detect the profiles of people who create fake news. Castillo et al. (2011) also use various characteristics of user profiles (e.g., registration age) to identify fake news. In addition to content and stylistic verification

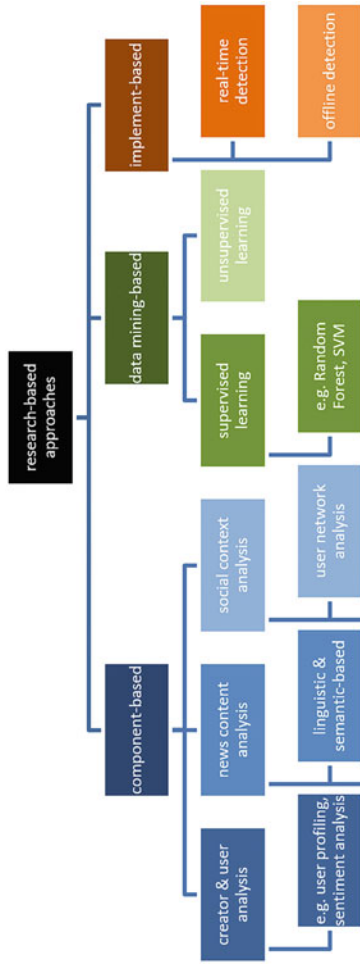

Fig. 7.2 Automatic detection methods of fake news (According to Zhang and Ghorbani 2020)

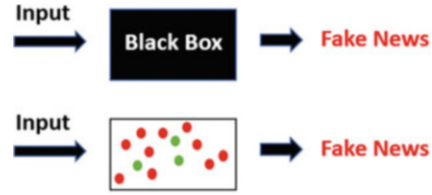

Fig. 7.3 Visualization of the black box (top) and white box approach (bottom)

mechanisms, there is the *propagation-based fake news detection* approach. This examines how news is propagated in social networks (Zhou et al. 2019).

7.2.2 Black Box Versus White Box

However, the aforementioned detection algorithms have a significant drawback: they are black-box based and accordingly do not provide end-users with an explanation of automated decision-making. Users can observe the input (e.g., a tweet) and the output (e.g., the marking of the tweet as fake news), but receive no information about what happens in between (e.g., why a tweet was marked as fake news). The counterpart of blackbox approaches is called whitebox approach. In this, the internal processes between input and output can be observed. In the context of fake news, whitebox approaches enable the traceability of indicators for false content. Accordingly, users here have access to all the necessary information to understand why the algorithm generated a specific output. A corresponding visualization is shown in Fig. 7.3.

In other contexts where machine learning is applied, the need for "interpretability, explainability and trustworthiness" is already highlighted and increasingly discussed (Conati et al. 2018, p. 24). Explainable machine learning sets out to build user trust in the results of systems (Ribeiro et al. 2016). So far, however, there are few approaches to fake news detection that use explainable machine learning. The approaches of Reis et al. (2019) and Yang et al. (2019) are worth mentioning.

Other whitebox approaches focus on user education with the aim of improving media literacy. Studies have shown that improved media literacy can have promising counteracting effects in dealing with fake news (Kahne and Bowyer 2017; Mihailidis and Viotty 2017). If the ability to autonomously evaluate online content is improved through whitebox approaches, this can reduce reactance and prevent the backfire effect. Nyhan and Reifler (2010) refer to the backfire effect as the emergence of anger and defiance when political content in particular contains a warning label. Users tend to believe the content all the more then, as they "perceive the correction as an illegal persuasion attempt" (Müller and Denner 2017, p. 17). Hartwig and Reuter (2019) designed a browser plugin that provides politically neutral and transparent cues about characteristics of a tweet on Twitter that indicate untrustworthy content. In a similar approach, Bhuiyan et al. (2018) present a browser plugin designed to help users on Twitter better assess the credibility of

news articles through nudging. Targeted questions (e.g., Does the post tell the whole story?) serve as a nudge to encourage users to think reflectively. Further, Fuhr et al. (2018) present an approach in which they label online texts in terms of, for example, facts and emotions, similar to nutritional information on food labels, to help readers make informed judgments. Instead of clear black-box or white-box approaches, platforms usually use combinations of different strategies to detect fake news. For example, Facebook offers users the possibility to report suspicious content and at the same time applies algorithms to detect and prioritize fake news, which are examined by independent fact checkers in the following (McNally and Bose 2018; Mosseri 2016).

7.3 Countermeasures to Support Users

A large body of academic work has already explored ways to automatically detect fake news. However, less attention has been paid to the next step, namely what to do when misinformation has finally been detected in social media. Has misinformation been successfully identified? Are there different approaches to deal with it in the following?

As misinformation spreads primarily through social media, platforms such as Facebook, Twitter and Instagram have begun to counteract it. Many of the approaches are directly visible to users and influence the experience on social networks. Facebook, in particular, has employed a number of practices as potential countermeasures since 2016 (Tene et al. 2018). For example, after the 2016 US election, Facebook began displaying warnings under controversial posts (Mosseri 2016). However, according to media reports, this feature was withdrawn after persistent criticism. Since then, Facebook has used more subtle techniques to limit the reach of controversial posts, such as reducing the post size, listing fact-check articles, and lowering the post ranking in the newsfeed (McNally and Bose 2018). These countermeasures appear to have roughly the desired effect of reducing the spread of fake news on the social network. Since their introduction in 2016, interaction with fake news on Facebook has been reduced by more than 50% (Allcott et al. 2019). In their paper, Kirchner and Reuter (2020) provide an overview of different social media techniques used. They further compare the effectiveness and user acceptance of different measures such as the display of warnings or related articles and the provision of additional information.

However, flagging and deleting false content may not be effective and sometimes even counterproductive. In contrast, many researchers see media literacy training as a promising strategy (Müller and Denner 2017; Stanoevska-slabeva 2017; Steinebach et al. 2020). Studies have shown that people with high media literacy are able to easily identify much of German-language fake news based on various factors such as text structure, as misinformation in the body of the text often has more than two spelling mistakes, consistent capitalization, or punctuation errors (cf. Steinebach et al. 2020). However, since most approaches to automatically detect and label fake news use black-box algorithms, and this is also the case with many widely used machine learning techniques, they usually cannot

clarify why they label certain content as fake news. Presenting users with a label can even lead to reactance if it does not match their own perception. This effect is generated by the so-called confirmation bias, which occurs when news is considered true precisely when it corresponds to one's ideology (Kim and Dennis 2018; Nickerson 1998; Pariser 2011).

Bode and Vraga (2015) investigated the possibility of combating misinformation with corrective information in the "Related Articles" section under the respective article. Researchers have previously shown that warnings about misinformation reduce its perceived accuracy (Ecker et al. 2010; Lewandowsky et al. 2012; Sally Chan et al. 2017), but these can also fail (Berinsky 2017; Nyhan and Reifler 2010; Nyhan et al. 2013). For example, Garrett and Weeks (2013) compared immediate versus delayed rectification for misformation. They found that immediate rectification had the most significant impact on perceived correctness. However, when misinformation confirms users' opinions, the potential for a backfire effect is greater (Kelly Garrett and Weeks 2013). Pennycook et al. (2018) show that a related phenomenon – repeated consumption of misinformation increases perceived Illusory Truth Effect accuracy – can also be applied to fake news on social media. In addition, they found that warnings can decrease the perceived accuracy of content. Pennycook et al. (2019) confirm the positive effect of such warnings. Using a Bayesian implied truth model, they argue that showing warning notifications for false news not only reduces belief in its accuracy, but also increases belief in the accuracy of news without an attached warning. Clayton et al. (2019) compare several types of warnings. In addition to specific warnings about false headlines, they also test a general warning without reference to a specific post. Facebook had displayed such a warning across users' newsfeed in April 2017 and May 2018, in which it warned against misinformation in general. In addition, they examine two different ways of phrasing specific warnings about headlines: "disputed" and "rated false". Their results show that general warnings have a minimal effect, but specific warnings have a significant effect. Thus, they confirm the findings of Pennycook et al. (2019). This group of researchers concluded that "rated false" warnings are significantly more effective than those labelled "disputed".

7.4 TrustyTweet: A Whitebox Approach to Assist Users in Dealing with Fake News

As shown in Sect.7.3, increasing media literacy is a promising strategy for dealing with fake news. By providing transparent and identifiable indicators of fake news, users can be supported in forming opinions about online content. In this context, it is important to differentiate between assistance systems that give neutral advice based on transparent indicators and systems that cause reactance in order to counteract a backfire effect. The use of a white box approach instead of a black box approach is an important step to reduce or prevent reactance.

In the following, the browser plugin TrustyTweet is presented, which aims to support users in dealing with fake news on Twitter by providing politically neutral, transparent and

Table 7.2 Potential indicators for fake news

Indicator	Example	Literature
Continuous capitalization	CONTINIOUS CAPITALIZATION	Steinebach et al. (2020); Wanas et al. (2008); Weerkamp and De Rijke (2008); Weimer et al. (2007)
Excessive use of punctuation	Excessive use of punctuation!!!	Morris et al. (2012);Wanas et al. (2008).
Wrong punctuation at the end of a sentence	Wrong punctuation at the end of the sentence!! 1	Morris et al. (2012); Weimer et al. (2007)
Excessive use of emoticons and especially attention-grabbing emoticons		Wanas et al. (2008); Weerkamp and De Rijke (2008)
The use of the standard profile screen		Morris et al. (2012)
Lack of official account verification, especially for celebrities		Morris et al. (2012)

intuitive advice (Hartwig and Reuter 2019). In particular, this approach aims to be a helpful assistant without leading to reactance. Users are thus not deprived of their own judgment. The aim is to bring about a learning effect regarding media literacy that makes the plugin redundant after prolonged use. In contrast to other approaches, TrustyTweets is therefore based on a white-box technology. The plugin was developed in a user-centered design process within the "design science" approach. Potential indicators of fake news were identified by weighing approaches that have already proven promising in scientific work. The focus is on heuristics that people intuitively and successfully use and that are easy to understand. However, it is important to emphasize that this approach cannot encompass all relevant indicators of fake news.

The following characteristics are used as potential indicators (Table 7.2):

TrustyTweet was developed for the Firefox web browser. Its main components are a text box that contains all the indicators detected in a tweet and serves as a warning notification, two different icons to indicate whether indicators have been detected in the tweet and, finally, another icon to access the settings that open in a popup window. Next to each indicator is a link to access general information about that indicator in a popup window. Moving the mouse over an indicator dynamically highlights the corresponding component in the tweet (see Fig. 7.4). This allows users to immediately see why a warning is displayed. The main icon of the plugin serves as a toggle button for the text box. Users can decide if they want to see all detected indicators next to the respective tweet or if they just want to see an icon and switch to the textbox if needed to see why the current warning is displayed. A key feature of TrustyTweet is the configuration popup. By using checkboxes, users can turn on and off individual indicators to investigate tweets. In this way, our plugin provides a stronger sense of autonomy and counters paternalism.

Fig. 7.4 Sample output from TrustyTweet. (Hartwig and Reuter 2019)

The usability and user experience of the plugin were evaluated in initial qualitative thinking aloud studies with a total of 27 participants. The support tool was largely rated as helpful and intuitive. Furthermore, the findings of our study provide indications for the following design implications for support tools in dealing with fake news:

1. Personalization to maintain personal autonomy: The configuration feature is important to increase autonomy and prevent reactance.
2. Support users by providing transparent and objective information: The indicators need detailed descriptions that make it clear why they are relevant for detecting fake news. According to our testers, it is of great importance that the descriptions are politically neutral and formulated in an objective way.
3. Clear mapping of alerts: Highlighting components of a tweet when hovering over it when a warning has been triggered has been deemed one of the most helpful plugin features and is indispensable to achieve a learning effect.
4. Personalized perceptibility: The toggle feature of the warnings was also positively received. Many participants liked the feature of displaying detailed text boxes only when needed and otherwise mainly paying attention to the color of the icon.
5. Minimizing false alarms: As in many other contexts (e.g. warning apps), it is very important to minimize false alarms, otherwise users might lose attention to the plugin or uninstall it before a learning effect has occurred. To improve the plugin in this respect, some respondents suggested the display of gradual warnings (for example in traffic light colors) as a possible alternative.

7.5 Conclusion and Outlook

Dealing with fake news is currently a major challenge for society and politics (cf. Granik and Mesyura 2017). Studies have shown that there is a great need for assistance systems to support social media users. So far, research has focused in particular on using machine learning algorithms to detect and label fake news. For example, Gupta et al. (2014) present a browser plugin that automatically assesses the truthfulness of content on Twitter. Other approaches (e.g., Fake News AI) also use machine learning. Still other approaches rely on whitelists and blacklists (e.g., B.S. Detector) to detect fake news. However, black-box methods run the risk of causing reactance, as they cannot give reasons for their fake news alerts.

In our eyes and following the opinion of other studies (Müller and Denner 2017; Stanoevska-slabeva 2017), improving individual media literacy is a central strategy in dealing with fake news. The initial empirical results of the conducted study show that our indicator-based white-box approach to support Twitter users in dealing with fake news is potentially promising if the following five design implications are considered: Personalizability to increase autonomy, transparent and objective information, unambiguity of warnings, personalized perceptibility, and minimization of false alarms. For future studies, a combination of automatic detection of fake news and subsequent use of TrustyTweet as a support measure is planned. Here, the advantages of both methods could be used: the transparent and easy-to-understand indicators and the accurate detection of black-box methods. A corresponding representative online experiment on the effectiveness of TrustyTweet in combination with automatic detection procedures as a supplement to the qualitative study conducted is being planned.

Acknowledgements Funded by the German Research Foundation (DFG) – SFB 1119 – 236615297 (CROSSING) as well as by the German Federal Ministry of Education and Research (BMBF) and the Hessian Ministry of Science and the Arts (HMWK) in the context of their joint funding for the National Research Center for Applied Cyber Security ATHENE.

This article is partly based on the article "*Fake News Perception in Germany: A Representative Study of People's Attitudes and Approaches to Counteract Disinformation*" (Reuter et al. 2019) and "*TrustyTweet: An Indicator-based Browser Plugin to Assist Users in Dealing with Fake News on Twitter*" (Hartwig and Reuter 2019). Moreover, it is partly based on the conference paper "*Countering Fake News: A Comparison of Possible Solutions Regarding User Acceptance and Effectiveness*" (Kirchner and Reuter 2020). We thank Jan Kirchner for his support.

References

Aisch G, Huang J, Kang C (2016) Dissecting the #PizzaGate conspiracy theories. New York times. https://www.nytimes.com/interactive/2016/12/10/business/media/pizzagate.html. Accessed on 18.04.2020

Allcott H, Gentzkow M (2017) Social media and fake news in the 2016 election. J Econ Perspect 31(2):211–236. https://doi.org/10.1257/jep.31.2.211

Allcott H, Gentzkow M, Yu C (2019) Trends in the diffusion of misinformation on social media. Res Politics 6(2):205316801984855. https://doi.org/10.1177/2053168019848554

Berinsky AJ (2017) Rumors and health care reform: experiments in political misinformation. Br J Polit Sci 47(2):241–262. https://doi.org/10.1017/S0007123415000186

Bhuiyan MM, Zhang K, Vick K, Horning MA, Mitra T (2018) Feed reflect: a tool for nudging users to assess news credibility on twitter. In: Companion of the 2018 ACM conference on computer supported cooperative work and social computing – CSCW '18, S 205–208. https://doi.org/10.1145/3272973.3274056

Bode L, Vraga EK (2015) In related news, that was wrong: the correction of misinformation through related stories functionality in social media. J Commun 65(4):619–638. https://doi.org/10.1111/jcom.12166

Bourgonje P, Moreno Schneider J, Rehm G (2018) From clickbait to fake news detection: an approach based on detecting the stance of headlines to articles. In: Proceedings of the 2017 EMNLP workshop: natural language processing meets journalism, S 84–89. https://doi.org/10.18653/v1/w17-4215

Breithut J (2020) Falschinformationen im Netz: so reagieren Facebook, Google und TikTok auf das Coronavirus. Spiegel online. https://www.spiegel.de/netzwelt/web/coronavirus-wie-facebook-google-und-tiktok-auf-falschinformationen-reagieren-a-6bc449fc-2450-4964-a675-7d6573316ad9. Accessed on 03.02.2020

Castillo C, Mendoza M, Poblete B (2011) Information credibility on twitter. In: Proceedings of the International Conference on World Wide Web, Hyderabad, S 675–684

Clayton K et al (2019) Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. Polit Behav 42:1073–1095. https://doi.org/10.1007/s11109-019-09533-0

Conati C, Porayska-Pomsta K, Mavrikis M (2018) AI in education needs interpretable machine learning: lessons from open learner modelling. In: Proceedings of 2018 ICML workshop on human interpretability in machine learning (WHI 2018). http://arxiv.org/abs/1807.00154. Accessed on 22.04.2021

Dutton WH, Fernandez L (2019) How susceptible are internet users? InterMedia 46(4). https://doi.org/10.2139/ssrn.3316768

Ecker UKH, Lewandowsky S, Tang DTW (2010) Explicit warnings reduce but do not eliminate the continued influence of misinformation. Mem Cogn 38(8):1087–1100. https://doi.org/10.3758/MC.38.8.1087

European Commission (2018) A multi-dimensional approach to disinformation. Report of the independent High Level Group on fake news and online disinformation (bd 2). https://doi.org/10.2759/0156

Fuhr N et al (2018) An information nutritional label for online documents. ACM SIGIR Forum 51(3):46–66. https://doi.org/10.1145/3190580.3190588

Granik M, Mesyura V (2017) Fake news detection using naive Bayes classifier. In: 2017 IEEE 1st Ukraine conference on electrical and computer engineering, UKRCON 2017 – proceedings, S 900–903. https://doi.org/10.1109/UKRCON.2017.8100379

Gravanis G, Vakali A, Diamantaras K, Karadais P (2019) Behind the cues: a benchmarking study for fake news detection. Expert Syst Appl 128:201–213. https://doi.org/10.1016/j.eswa.2019.03.036

Grinberg N, Joseph K, Friedland L, Swire-Thompson B, Lazer D (2019) Political science: fake news on twitter during the 2016 U.S. presidential election. Science 363(6425):374–378. https://doi.org/10.1126/science.aau2706

Gupta A, Kumaraguru P, Castillo C, Meier P (2014) TweetCred: real-time credibility assessment of content on twitter. In: International conference on social informatics, S 228–243. http://arxiv.org/abs/1405.5490. Accessed on 22.04.2021

Hanselowski A, Stab C, Schulz C, Li Z, Gurevych I (2019a) A richly annotated corpus for different tasks in automated fact-checking. In: proceedings of the 23rd conference on computational natural language processing, S 493–503. https://doi.org/10.18653/v1/k19-1046

Hanselowski A et al (2019b) UKP-Athene: multi-sentence textual entailment for claim verification. In: proceedings of the first workshop on fact extraction and verification (FEVER), S 103–108. https://doi.org/10.18653/v1/w18-5516

Hartwig K, Reuter C (2019) TrustyTweet: an indicator-based browser-plugin to assist users in dealing with fake news on twitter. In: proceedings of the international conference on Wirtschaftsinformatik (WI). http://www.peasec.de/paper/2019/2019_HartwigReuter_TrustyTweet_WI.pdf. Accessed on 18.04.2020

Jin Z, Cao J, Zhang Y, Luo J (2016) News verification by exBploiting conflicting social viewpoints in microblogs. In: 30th AAAI conference on Artificial Intelligence, AAAI 2016, Phoenix, S 2972–2978

Kahne J, Bowyer B (2017) Educating for democracy in a partisan age: confronting the challenges of motivated reasoning and misinformation. Am Educ Res J 54(1):3–34. https://doi.org/10.3102/0002831216679817

Kaufhold M, Rupp N, Reuter C, Habdank M (2020) Mitigating information overload in social media during conflicts and crises: design and evaluation of a cross-platform alerting system. Behav Inform Technol 39(3):319–342

Kelly Garrett R, Weeks BE (2013) The promise and peril of real-time corrections to political misperceptions. Proceedings of the ACM conference on Computer Supported Cooperative Work, CSCW, S, In, pp 1047–1057. https://doi.org/10.1145/2441776.2441895

Kim A, Dennis A (2018) Says who?: how news presentation format influences perceived believability and the engagement level of social media users. In: proceedings of the 51st Hawaii international conference on system sciences. https://doi.org/10.24251/hicss.2018.497

Kirchner J, Reuter C (2020) Countering fake news: a comparison of possible solutions regarding user acceptance and effectiveness. In: proceedings of the ACM: human computer interaction (PACM): computer-supported cooperative work and social computing, ACM, Austin, USA

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

Long Y, Lu Q, Xiang R, Li M, Huang C-R (2017) Fake news detection through multi-perspective speaker profiles, Bd 2, 8. Aufl. In: Proceedings of the eighth international joint conference on Natural Language Processing, Taipei, S 252–256

McNally M, Bose L (2018) Combating false news in the Facebook news feed: fighting abuse @scale. https://atscaleconference.com/events/fighting-abuse-scale/. Accessed on 24.01.2020

Mihailidis P, Viotty S (2017) Spreadable spectacle in digital culture: civic expression, fake news, and the role of media literacies in "post-fact" society. Am Behav Sci 61(4):441–454. https://doi.org/10.1177/0002764217701217

Monti F, Frasca F, Eynard D, Mannion D, Bronstein MM (2019) Fake news detection on social media using geometric deep learning. [Preprint]

Morris MR, Counts S, Roseway A, Hoff A, Schwarz J (2012) Tweeting is believing? Understanding microblog credibility perceptions. Proceedings of the ACM conference on Computer Supported Cooperative Work, CSCW, S, In, pp 441–450. https://doi.org/10.1145/2145204.2145274

Mosseri A (2016) Addressing hoaxes and fake news. https://about.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/. Accessed on 24.01.2020

Müller P, Denner N (2017) Was tun gegen „Fake News"? Friedrich Naumann Stiftung Für die Freiheit, Bonn

Nickerson RS (1998) Confirmation bias: a ubiquitous phenomenon in many guises. Rev Gen Psychol 2:175–220

Nyhan B, Reifler J (2010) When corrections fail: the persistence of political misperceptions. Polit Behav 32(2):303–330. https://doi.org/10.1007/s11109-010-9112-2

Nyhan B, Reifler J, Ubel PA (2013) The hazards of correcting myths about health care reform. Med Care 51(2):127–132. https://doi.org/10.1097/MLR.0b013e318279486b

Pariser E (2011) The filter bubble: how the new personalized web is changing what we read and how we think. Penguin, London

Pennycook G, Cannon TD, Rand DG (2018) Prior exposure increases perceived accuracy of fake news. J Exp Psychol Gen 147(12):1865–1880. https://doi.org/10.1037/xge0000465

Pennycook G, Bear A, Collins E (2019) The implied truth effect: attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. In: management science. http://www.mnsc.2019.3478.pdf. Accessed on 18.04.2020

Pérez-Rosas V, Kleinberg B, Lefevre A, Mihal R (2017) Automatic detection of fake news. In: proceedings of the 27th international conference on computational linguistics. https://www.aclweb.org/anthology/C18-1287. Accessed on 22.04.2021

Potthast M, Kiesel J, Reinartz K, Bevendorff J, Stein B (2018) A stylometric inquiry into hyperpartisan and fake news. In: ACL 2018 – 56th annual meeting of the Association for Computational Linguistics, proceedings of the conference (Long papers) Vol. 1, S 231–240. https://doi.org/10.18653/v1/p18-1022

Rapoza K (2017) Can "fake news" impact the stock market? In: Forbes. https://www.forbes.com/sites/kenrapoza/2017/02/26/can-fake-news-impact-the-stock-market/#33dc99c02fac. Accessed on 24.01.2020

Rashkin H, Choi E, Jang JY, Volkova S, Choi Y (2017) Truth of varying shades: analyzing language in fake news and political fact-checking. In: EMNLP 2017 – conference on empirical methods in natural language processing, proceedings, S 2931–2937. https://doi.org/10.18653/v1/d17-1317

Reis JCS, Correia A, Murai F, Veloso A, Benevenuto F (2019) Explainable machine learning for fake news detection. In: WebSci 2019 – proceedings of the 11th ACM conference on web science, S. Association for Computing Machinery, Inc, pp 17–26. https://doi.org/10.1145/3292522.3326027

Reuter C, Hartwig K, Kirchner J, Schlegel N (2019) Fake news perception in Germany: a representative study of people's attitudes and approaches to counteract disinformation. In: proceedings of the international conference on Wirtschaftsinformatik. http://www.peasec.de/paper/2019/2019_ReuterHartwigKirchnerSchlegel_FakeNewsPerceptionGermany_WI.pdf. Accessed on 18.04.2020

Ribeiro MT, Singh S, Guestrin C (2016) „Why should i trust you?" explaining the predictions of any classifier. In: proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining, S 1135–1144. Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

Ruchansky N, Seo S, Liu Y (2017) CSI: a hybrid deep model for fake news detection. In: international conference on information and knowledge management, proceedings, S 797–806. Association for Computing Machinery. https://doi.org/10.1145/3132847.3132877

Sally Chan M, Jones CR, Hall Jamieson K, Albarraci D (2017) Debunking: a meta-analysis of the psychological efficacy of messages countering misinformation. Psychol Sci 28(11):1531–1546. https://doi.org/10.1177/0956797617714579

Sängerlaub A (2017) Verzerrte Realitäten – Die Wahrnehmung von „Fake News" im Schatten der USA und der Bundestagswahl. Stiftung Neue Verantwortung, Berlin. https://www.stiftung-nv.de/sites/default/files/fake_news_im_schatten_der_usa_und_der_bundestagswahl.pdf. Accessed on 18.04.2020

Sethi RJ (2017) Crowdsourcing the verification of fake news and alternative facts. In: HT 2017 – proceedings of the 28th ACM conference on hypertext and social media, S. Association for Computing Machinery, Inc, pp 315–316. https://doi.org/10.1145/3078714.3078746

Shu K, Bernard HR, Liu H (2019a) Studying fake news via network analysis: detection and mitigation. In: Emerging research challenges and opportunities in computational social network analysis and mining, S 43–65. https://doi.org/10.1007/978-3-319-94105-9_3

Shu K, Wang S, Liu H (2019b) Beyond news contents: the role of social context for fake news detection. In: WSDM 2019 – proceedings of the 12th ACM international conference on web search and data mining, S. Association for Computing Machinery, Inc., pp 312–320. https://doi.org/10.1145/3289600.3290994

Stanoevska-slabeva K (2017) Teaching social media literacy with storytelling and social media curation. In: twenty-third Americas conference on information systems, S 1. https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1524&context=amcis2017. Accessed on 18.04.2020

Steinebach M, Bader K, Rinsdorf L, Krämer N, Roßnagel A (2020) Desinformation aufdecken und bekämpfen: Interdisziplinäre Ansätze gegen Desinformationskampagnen und für Meinungspluralität, Bd 45, 1. Aufl. Nomos Verlagsgesellschaft mbH & Co. KG, Baden-Baden. https://doi.org/10.5771/9783748904816

Tacchini E, Ballarin G, Della Vedova ML, Moret S, de Alfaro L (2017) Some like it hoax: automated fake news detection in social networks. In: CEUR workshop proceedings (Vol. 1960). https://developers.facebook.com/docs/graph-api. Accessed on 22.04.2021

Tene O, Polonetsky J, Sadeghi A-R (2018) Five freedoms for the momodeus. IEEE Secur Priv 16(3): 15–17. https://ieeexplore.ieee.org/abstract/document/8395137/. Accessed on 22.04.2021

Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Science 359:1146–1151

Wanas N, El-Saban M, Ashour H, Ammar W (2008) Automatic scoring of online discussion posts. In: International conference on Information and Knowledge Management, Proceedings, S 19–25. https://doi.org/10.1145/1458527.1458534

Weerkamp W, De Rijke M (2008) Credibility improves topical blog post retrieval. In: ACL-08: HLT – 46th annual meeting of the Association for Computational Linguistics: human language technologies, proceedings of the conference, Columbus, S 923–931

Weimer M, Gurevych I, Mühlhäuser M (2007) Automatically assessing the post quality in online discussions on software. In: proceedings of the 45th annual meeting of the Association for Computational Linguistics Companion Volume Proceedings of the demo and poster sessions, S 125–128. Association for Computational Linguistics. https://doi.org/10.3115/1557769.1557806

Wu L, Liu H (2018) Tracing fake-news footprints: characterizing social media messages by how they propagate. In: WSDM 2018 – proceedings of the 11th ACM international conference on web search and data mining, S. Association for Computing Machinery, Inc, pp 637–645. https://doi.org/10.1145/3159652.3159677

Yang F et al (2019) XFake: explainable fake news detector with visualizations. In: The Web Conference 2019 – proceedings of the World Wide Web Conference, WWW 2019, S 3600–3604. Association for Computing Machinery, Inc https://doi.org/10.1145/3308558.3314119

Zhang X, Ghorbani AA (2020) An overview of online fake news: characterization, detection, and discussion. Inf Process Manag 57(2):102025. https://doi.org/10.1016/j.ipm.2019.03.004

Zhou X, Jain A, Phoha VV, Zafarani R (2019) Fake news early detection: a theory-driven model. Digit threats res Pract. http://arxiv.org/abs/1904.11679. Accessed on 22.04.2021

Katrin Hartwig, M.Sc., co-authored the paper "Countering Fake News Technically – Detection and Countermeasure Approaches to Support Users" with Christian Reuter. She is a research associate at the Chair of Science and Technology for Peace and Security (PEASEC) at TU Darmstadt and works in the fields of human-computer interaction, disinformation in social media and usable security.

Christian Reuter, Prof. Dr., co-authored the paper "Countering Fake News Technically – Detection and Countermeasure Approaches to Support Users" with Katrin Hartwig. He holds the Chair of Science and Technology for Peace and Security (PEASEC) at TU Darmstadt and works in the fields of security-critical human-computer interaction, IT for peace and security, and resilient IT-based (critical) infrastructures.