TrustyTweet: An Indicator-based Browser-Plugin to Assist Users in Dealing with Fake News on Twitter

Katrin Hartwig and Christian Reuter

Technische Universität Darmstadt,
Science and Technology for Peace and Security (PEASEC), Germany
katrin.hartwig@stud.tu-darmstadt.de
reuter@peasec.tu-darmstadt.de

Abstract. The importance of dealing with fake news on social media has increased both in political and social contexts. While existing studies focus mainly on how to detect and label fake news, approaches to assist users in making their own assessments are largely missing. This article presents a study on how Twitter-users' assessments can be supported by an indicator-based white-box approach. First, we gathered potential indicators for fake news that have proven to be promising in previous studies and that fit our idea of a white-box approach. Based on those indicators we then designed and implemented the browser-plugin *TrusyTweet*, which assists users on Twitter in assessing tweets by showing politically neutral and intuitive warnings without creating reactance. Finally, we suggest the findings of our evaluations with a total of 27 participants which lead to further design implications for approaches to assist users in dealing with fake news.

Keywords: Fake News, Social Media, Twitter, Plugin

1 Introduction

Fake news can be defined as "*news articles that are intentionally and verifiably false and could mislead readers*" [1]. Recently, the term has gained popularity, especially in discussions concerning the political context. The U.S. presidential election in 2016, as well as the German parliamentary election in 2017 among others, showed a great perceived significance of fake news for the society. Although studies have shown that there were no impacts on the election outcomes [1], the society fears the effect of fake news in social media. Our previous representative study on the perception of fake news in Germany revealed that 84 % of the citizens agree with fake news posing a threat [2]. Those concerns are not groundless as fake news can indeed have serious consequences. For example, in 2013 the official Twitter account of Associated Press (AP) was hacked. In consequence, the stocks experienced a temporary loss of $130 billion [3]. Furthermore, fake news can be relevant in the context of peace and political propaganda [4]. Thus, finding adequate strategies to counteract the negative effects of fake news, especially in social networks, is of high interest. Examining fake news in online information is highly relevant in the IS research field [5]. Several studies have already

14th International Conference on Wirtschaftsinformatik,
February 24-27, 2019, Siegen, Germany

shown that labeling and deleting fake contents is not effective and sometimes counterproductive. Instead, scientists argue that the training of media literacy is a promising strategy [6], [7]. (Media) literacy is defined as the ability to access, analyze, evaluate and create messages in a variety of forms [8]. However, most approaches concentrate on black-box algorithms to automatically detect and label fake news. In black-box approaches one can observe the input (in our case e.g. a tweet) and the output (here e.g. the label as "fake") but there is no information about what happens in between (e.g. why the tweet was labeled as "fake"). The counterpart is called white-box approach, where internals can be reviewed. In our context, white-box approaches facilitate the comprehension of reasons that indicate fake content, so that the user has all necessary information to understand why an algorithm has a specific output.

The objective of this article is to examine how users on Twitter could be supported in dealing with fake news by a white-box-based browser-plugin. Our research questions are: How can we provide a transparent, politically neutral and objective assisting tool for users of social media? Moreover, does a white-box approach counteract reactance and encourage a learning effect? The article is organized as follows: Section 2 presents related work on assisting tools to counteract fake news. Section 3 presents our research approach of *design science* [9], which focuses on the design of an artifact for a relevant problem and rigorous evaluation methods. In section 4 we propose the concept of *TrustyTweet*, a white-box plugin for users on Twitter whose evaluation will be presented in section 5. Finally, we discuss the potential scientific contributions and limitations of our approach in section 6.

2 Related Work & Research Gap

While the effect of fake news has proven to be significant in specific cases and the debates in politics and society continue, several approaches try to find answers on how to counteract fake news. Recently, many studies have been conducted to detect and label fake news. Viviani and Pasi present a survey on how approaches automatically assess credibility in online review sites, microblogs and sources of online health information [10]. Rubin presents the state-of-the-art technologies on fake news detection, divided into linguistic and network approaches [11]. Both studies show that most of the approaches use machine learning techniques to identify fake contents. Despite machine learning algorithms, blacklists and whitelists of websites are commonly used. In the following list, an excerpt of existing browser-plugins and smartphone apps is presented which we found searching the Google Play Store, browser add-on sections and scientific contributions. Relevant associated characteristics were extracted from the official descriptions and are given in **Table 1**.

- **TweetCred**: Browser-plugin with a semi-supervised ranking model using SVM-rank to assess credibility in tweets. Displays a credibility score from 1 to 7. [12]
- **B.S. Detector**: Searches all links on a given webpage for references to unreliable sources, checking against a manually compiled list of domains. (http://bsdetector.tech)

- **Fake News Detector**: Allows users to tag news stories on Facebook and Twitter. Tags will be stored in a database and used by an AI to learn. In the future contents will be highlighted based on user input and the AI. (https://fakenewsdetector.org)
- **Fake News AI**: Uses a neural network to analyze writing, sophistication, site popularity, content and many more. (http://www.fakenewsai.com)
- **Fake News Check**: A smartphone app that does not detect fake news automatically but causes to reflect by asking 19 relevant questions which the user needs to answer to receive feedback. The app was developed for students to train media literacy. (https://www.neue-wege-des-lernens.de/projekte/fake-news-check)

Table 1. Limitations of existing approaches

Name of application	*Binary labels*	*Gives transparent reasons*	*Provides learning effect*	*Uses a database (black- or whitelist)*	*Based on training data*	*Big effort for users*
TweetCred					X	
B.S. Detector	X			X		
Fake News Detector				X	X	X
Fake News AI	X				X	
Fake News Check		X	X			X

Rehm states that fully automatic technologies are partly suitable to support the user in dealing with fake news but cannot take over all necessary tasks. Additionally, approaches have to be based on human intelligence [13]. The stand-alone usage of blacklists and whitelists works only for websites or other texts that contain links to URLs included in one of those lists. Since the online environment is much more complex, manually compiling lists with reliable or unreliable sources does not lead to sufficient results [10]. Despite the absence of gold standard datasets to train the classifiers [10], machine learning algorithms have another major flaw if used to assist the user. As machine learning techniques are black-box procedures, they cannot reason why they label a content as fake. Showing the user a label, it might even create reactance if it does not fit his or her own perception. That effect is caused by the *Confirmation Bias* due to which messages are particularly considered true if they fit the own ideology [14–16]. None of the existing approaches gives transparent reasons or encourages a learning effect while leading to little effort for the user.

Studies have shown that a promising way to counteract fake news is increasing media literacy [17, 18]. Improving the capacity to evaluate online contents as autonomously as possible using white-box instead of black-box approaches can minimize reactance and prevent the *Backfire Effect*. While several approaches use machine learning techniques to label contents as fake or not fake, there are no approaches that work with a white-box technique despite "Fake News Check". This application, though, involves a big effort for the user as he must manually go through 19 questions for every text he wants to check before receiving a feedback. The approach itself does not include an automatic check to detect fake news. To date, no approach to detect fake news is based on the ideas of media literacy or white-box methods.

3 Research Design

Keeping in mind the detected limitations of machine learning-based approaches, for instance regarding created reactance, our intention is to offer a first survey on how white-box approaches can help to counteract fake news in social media. Our aim is to find answers to the following research questions: *How can we provide a transparent, politically neutral and objective assisting tool for users of social media? Furthermore, does a white-box approach counteract reactance and encourage a learning effect?* To encounter the lack of empirical findings regarding white-box approaches in the given context, we present a browser-plugin for Twitter which has been developed and evaluated in an iterative process. The plugin focuses on Twitter as it is a popular platform for breaking news with a high relevance of fake news, for instance in emergency situations but also in political campaigns [19]. As a popular communication channel, it is commonly used in scientific studies to examine social media from various perspectives (e.g. [20 21]). We used the *design science* approach [9] which focuses on the design of an artifact for a relevant problem and rigorous evaluation methods. In our case, the artifacts will be versions of our plugin and evaluations will take place in form of thinking-aloud studies. The method applies five steps, namely (a) achieving a problem awareness, (b) suggesting solutions, (c) development of solutions, (d) evaluation and finally (e) conclusion. The *design science* approach has proven to be an appropriate method to create new and innovative artifacts [9].

4 Concept of *TrustyTweet*

Instead of labeling and deleting, acquiring a high standard of media literacy is considered to be a promising approach in combating the impact of fake news. Given a number of transparent and identifiable indicators for fake news, the user of social media can be supported in forming an opinion about online content. In that context, it is crucial to differentiate between assistance systems that give neutral hints based on transparent indicators to train media literacy and systems that create reactance. Müller and Denner indicate that warning messages might lead to a *Backfire Effect* [6]. Especially in political contexts, users might rate the warning message as an illegal attempt to persuade the user which can result in believing in the content even more. Using a white-box instead of a black-box procedure is an important step to prevent or minimize reactance.

Our aim is to support the user in dealing with fake news in tweets and to increase his media literacy. We present *TrustyTweet*, a browser-plugin that intends to support users on Twitter in dealing with fake news by giving politically neutral, transparent and intuitive hints. This approach particularly aims to be a helpful assistant without creating reactance. The user continues to be in the power of his own assessments. We intend to create a learning effect regarding media literacy to make the plugin redundant after a longer regular usage. Therefore, different from other approaches, *TrustyTweet* is based on white-box technology. The plugin was developed in a user-centered design process using the design science approach. We identified potential indicators for fake news by

considering what has already been proven in studies to be successful. The focus is on heuristics that are used intuitively and successfully by humans and are easily comprehensible for everyone. In several qualitative thinking-aloud studies we evaluated the perceived helpfulness, the users' perceived autonomy and usability of the plugin on Twitter.

4.1 Identification of Indicators

Since the plugin aims to be a white-box approach it intends to show transparent warnings which the user can comprehend at any time, regardless of his level of media literacy. Morris et al. [22] found that users assess contents especially using features that are visible at a glance. In our approach, we intend to follow that idea. Potential indicators that fit our intentions and that largely have already been proven in studies to be promising in indicating fake news are the following:

- Consecutive capitalization [23-25]
- Excessive usage of punctuation (e.g. "*!!!*") [22], [23]
- Wrong punctuation a the end of sentences [22], [24]
- Excessive usage of emoticons [23], [25] and attention-grabbing emoticons
- Default account image [22]
- The absence of an official account verification seal (especially for celebrities) [22]

The potential indicators were assessed in a first thinking-aloud pre-study with six participants aged 23 to 28 (4 female, 2 male, all university students) in an average duration of 33 minutes. All six indicators have proven to be easily comprehensible for our test subjects. Furthermore, the detection algorithm for each indicator has an acceptable runtime to support users on Twitter dynamically and in real time. It is vital to clarify that our approach does not claim to comprehend all relevant indicators for fake news. The mentioned group of indicators includes those that are fitting our white-box idea since they are easily comprehensible.

4.2 Underlying Technology and Components of the Plugin

TrustyTweet was developed for the Firefox browser and uses jQuery and Semantic UI. Its main components are a textbox containing all indicators detected in a specific tweet which serves as a warning, two distinct icons to report if indicators have been detected in the specific tweet or not and an icon to open the settings in a pop-up window. The indicators were detected by searching the DOM tree of Twitter. Next to each indicator, there is a link to open more generic information about the indicator in a pop-up window (see **Figure 3**). When hovering the mouse over an indicator, the underlying component of the tweet is being highlighted dynamically (see **Figure 1**). Hence, the user can see immediately why the warning is being displayed. The main icon of the plugin serves as a toggle button for the textbox. The user can decide if he wants to see all detected indicators next to each tweet or if he prefers to see only the icon and toggle the textbox whenever he is interested in why the warning is being displayed. Additionally, it

guarantees that other contents like "*Who to follow*" do not remain hidden. A central feature of *TrustyTweet* is the configuration pop-up (see **Figure 2**). Using checkboxes, the user can switch the examination regarding specific indicators on and off. Hence, the plugin intends to build a stronger sense of autonomy and to counteract paternalism.

Figure 1. Exemplary output of plugin for four tweets

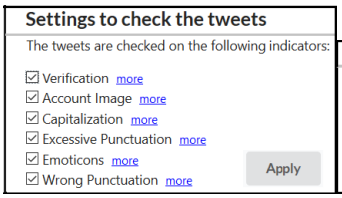

Figure 2. Pop-up with settings

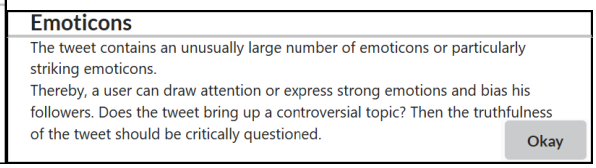

Figure 3. Pop-up with additional information

5 Evaluation

5.1 Methodology

Using the design science approach, we iteratively applied five steps to achieve a problem awareness, to suggest solutions in form of potential plugin-designs, to implement those solutions, to thoroughly evaluate them and to finally draw a conclusion. The iteratively conducted evaluations were based on the thinking-aloud method in which the user explains why he carries out which activity, which information is incomprehensible or does not meet his expectations and what he likes or dislikes. The audio and video material was recorded using the Xbox DVR-tool. While using *TrustyTweet* on Twitter, the subjects were asked to execute usability tasks and answer open questions. The average durations were 33 minutes in the first pre-study and 11

minutes in the second and main study. The participants were informed beforehand that there is no "*right*" or "*wrong*" for answering the questions. To guarantee the same conditions for all subjects in the first subtest, tweets were generated by a test account by which the subtest was performed. Additionally, a second subtest was performed on real-time tweets of a politician and a German news page on Twitter to receive an impression of the usage of the plugin in a realistic environment. After each study, the detected flaws were patched and suggested improvements were implemented to receive better results in the next iteration. While the tasks in the first formative study focus mainly on the comprehensibility of the suggested indicators and gather ideas on how to increase the perceived autonomy, the second study includes an examination of the central configuration feature and the realization of dialogue design principles. Since interactions are a central component of our plugin, we want to gather information about the fulfillment of dialogue principles as conformity with user expectations and self-descriptiveness. The third study intends to examine to what extent a well-usable version of the plugin supports the user in dealing with fake news, including a summative evaluation of usability. Continuing the iterative process our aim is to understand to what extent users feel autonomous or patronized during the usage of our white-box-based plugin. Furthermore, we intend to examine the perceived helpfulness of the plugin. Therefore, the test subjects were asked to perform several tasks (e.g. Open the configurations of the plugin. Check the tweets on "*capitalization*" and "*emoticons*" only.) and answer specific questions (e.g. To what extent do you feel patronized or autonomous when assessing the tweets? / How do you like the plugin contentwise? Is it helpful or is it obstructive?).

Characteristics of Study Participants. In the first pre-study, a number of six participants (4 female, 2 male) took part, in the second pre-study a number of five (2 female, 3 male) and in the main study a number of 16 participants (7 female, 9 male). The participants' age ranged from 23 to 28 years in the first and second study and from 21 to 34 years in the main study. The majority of the test subjects were university students (19 out of 27 in total) due to their good accessibility for scientific studies and their relevance as potential Twitter-users. The remaining eight participants stated to be employees. In the first pre-study, three out of six participants and in the second pre-study three out of five participants stated that they have a Twitter account or that they had an account in the past. In the main study, it applied to half of the test subjects. Participants in all studies that stated to have never had a Twitter account were introduced to the central aspects and components of Twitter and its tweets before they started completing the tasks.

Analysis. Following the standard proceeding for thinking-aloud tests according to van Someren, Barnard & Sandberg [26] we examined the obtained qualitative data of all thinking-aloud studies by reviewing the video and audio material, transcribing all important statements and assigning the statements to their associated tasks and actions. The statements were then clustered thematically and gathered for all test subjects. Eight categories were developed inductively from the data (helpfulness, autonomy, additional information, configuration, toggle-feature, mouseover-feature, salience, other). Each statement was assigned to a category by looking for keywords (e.g. "patronized": autonomy), considering the context of tasks. Hence, conclusions were drawn from the

various categories. The most noteworthy contributions of the main study will be presented in the following chapter.

5.2 Empirical Results

Perceived Helpfulness. When asked about helpfulness or obstructiveness of the plugin, 13 out of 16 participants regarded it to be a helpful tool. Most participants appreciated particularly its transparent nature and the simple visual feedback as well as the possibility to toggle the textbox while still getting a feedback from the icon: *"I like it a lot. You must keep thinking for yourself, but the plugin makes it easier and things attract your attention faster. It says: Attention! Here it would be wiser to think about the tweet again" (E12 #00:09:28).*

On the other hand, three participants argued that they personally do not need the plugin and therefore could not yet see the added value. They pointed out, that the warnings were based on very simple indicators, which they were able to detect by themselves: *"I think I do not need it. It is very interesting, that the displayed warnings are exactly the things I use as a search filter in my head when I read texts." (E15 #00:08:53).* One participant was concerned about the plugin showing too many false alarms, for example when warning of non-celebrity users that are not verified: *"If after every storm you warn against there are only three drops of rain, I will eventually not pay attention to it anymore. Therefore, it might be better to raise the threshold or to show graduated warnings." (E27 #00:07:24).*

More positively, one participant highlighted the desired learning effect of the plugin: *"I can imagine that it is very good to learn what you have to pay attention to. At some point when you have enough practice, you have taken on the same policies." (E14 #00:06:14).* Additionally, the participants had some interesting ideas to improve the plugin. For instance, it would be a helpful feature to display a link to the scientific sources of the chosen indicators. This is in the spirit of our white-box approach and would enhance transparency and objectivity. Furthermore, as the plugin does not include checks on videos and images, that should be pointed out to the user.

Perceived Autonomy. When we asked the participants to what extent they felt autonomous or patronized when assessing the tweets, all 16 participants regarded not to feel patronized at all. They highlighted the neutrally phrased additional information and the fact, that the plugin does not decide if a tweet contains fake news: *"I did not find it patronizing at all, especially because the explanations are written very neutrally. (...) You still must keep thinking for yourself. The plugin says there might be an indicator, but it does not have to be fake." (E14 #00:06:58).* The perceived autonomy was also enhanced by the configuration feature: *"When punctuation is not a criterion for me, I can just switch it off" (E27 #00:09:07).*

Plugin Features: Usability and Layout. All but one participant managed to interact with the additional information pop-up intuitively and very fast. Five participants noted explicitly that the additional information was helpful and necessary to understand the indicators completely. The configuration-feature has proven to be a substantially important aspect to enhance the perceived autonomy already in the two first studies using low-fidelity prototypes. The fully implemented version was used effectively and

intuitively by all 16 participants of the third study: *"Wow, that is easy! I do not have to think at all" (E21 #00:02:25).*
12 out of 16 participants managed to toggle the textbox containing the detected indicators straightaway. On the other hand, four participants struggled initially to find the correct button and tried a Twitter-internal button before toggling the textbox successfully. Most participants valued the feature since it can make the feedback more compact. The highlighting-feature was appreciated a lot by all participants. Some participants even said it was the most helpful component of the plugin as it helps the user to comprehend all warnings. Hence, the feature is central to our white-box approach. Using the highlighting-feature, all participants were able to match a specific warning to the correct referring part of the tweet successfully. All but two participants stated to like the layout of the plugin. While eleven out of 16 participants said it was noticeable enough, four were undecided and one found it was too noticeable. To avoid misunderstandings, participants suggested to add a mouseover-effect to the icon which appears when no implemented indicator was detected in a specific tweet.

5.3 Concluding Design Implications

Considering the presented results regarding perceived helpfulness, autonomy and usability of the plugin, we present five design implications to enhance the value of an indicator-based white-box approach to support users on Twitter in dealing with fake news. Those implications were extracted from statements that were mentioned particularly often or highlighted as very crucial by the participants. They support the view of existing studies, for instance by highlighting the relevance of transparent information (e.g. [6]) and the minimization of false alarms [27]. Moreover, they expand existing knowledge by a first scientific contribution on how to successfully develop an indicator-based white-box approach in that specific context.
Personalization to enhance autonomy. The configuration-feature is substantially important to enhance the perceived autonomy of the users and to prevent reactance. Our test subjects endorsed the possibility of deciding for themselves, on which indicators the tweets should be checked.
Assisting with transparent and objective information. The indicators need a detailed description that explains why they are relevant with regards to fake news. According to our test subjects, it is crucial, that the descriptions are formulated in a politically neutral and objective way. Adding a link to the associated scientific study can increase credibility in the spirit of our white-box approach. Warnings should always clarify that are assisting but do not replace the users' own assessment. Furthermore, the users must be aware of what functionalities the plugin does not include (e.g. our plugin does not examine images and videos).
Unambiguousness of warnings. Highlighting parts of the tweet in a mouseover-effect concerning a displayed warning was rated as one of the most helpful features of the plugin. Matching specific warnings to the correct referred part of the tweet is central to enhance the desired learning effect.
Personalized noticeability. The toggle-feature of the warnings was rated positively by the test subjects. Since the icon still gives a simple visible feedback, it is a pleasant way

of making the plugin more compact and adjusting it to the users' preferences concerning noticeability.

Minimizing false alarms. Minimizing false positives is crucial to prevent users from not paying attention anymore or uninstalling the plugin before having achieved a learning effect. Therefore, the threshold of warnings should not be too low. Some participants see advantages in showing graduated warnings in different colors.

6 Discussion & Conclusion

Dealing with fake news has proven to be one of the big current challenges in society and politics. Studies have shown that there is a need for assisting tools to support users on social media. There has been previous research on using machine learning algorithms to detect and label fake news. For example, Gupta et al. [12] present a browser-plugin to automatically assess the credibility of contents on Twitter. Further approaches (e.g. Fake News AI) use machine learning techniques as well. Other approaches are based on whitelists or blacklists (e.g. B.S. Detector) to detect fake news. The usage of black-box approaches though is not able to give reasons for its decisions and therefore, it runs the risk of creating reactance. In our eyes and according to other studies ([6], [7]), improving media literacy is a crucial strategy to help users dealing with fake news. Therefore, white-box approaches are necessary. However, all presented plugins, applications and approaches are based on black-box methods. Although the smartphone application Fake News Check can give transparent reasons on why contents might be fake, it does not automatically check for indicators and it comes with a big effort for the user.

Our scientific contribution is to theoretically explore the potential of an indicator-based white-box approach to assist users on Twitter and more practically to design, implement and evaluate a consistent browser-plugin as an artifact regarding to the design science approach. The plugin includes a warning with regards to six easily comprehensible and politically neutral indicators for fake news, further information about every indicator and a configuration-feature to support personalization. To answer our first research question (*How can we provide a transparent, politically neutral and objective assisting tool for users of social media?*), the empirical findings in Section 5.3 reveal that our indicator-based white-box approach to support users on Twitter in dealing with fake news can be considered suitable, applying the following five design implications: personalization to enhance autonomy, transparent and objective information, unambiguousness of warnings, personalized noticeability and minimization of false alarms. Moreover, we intended to answer the second research question: *Does a white-box approach counteract reactance and encourage a learning effect?* Our study shows that our white-box approach is a promising way to support users on social media without creating reactance but encouraging a learning effect and can therefore be considered a suitable alternative to black-box approaches.

Following our concept of design science, we intend to evaluate the newly suggested features (e.g. graduated warnings) in the future. Moreover, we intend to integrate further relevant user groups in our evaluation. In addition to our qualitative studies, a

quantitative study is desirable to guarantee an evaluation on a larger scale. On the other hand, it would be interesting to examine if there is a beneficial way to combine our white-box tool with the features of a machine learning approach to receive the advantages of both methods, namely the transparent and easily comprehensible indicators of our approach and the accurate classifications of black-box approaches.

References

1. Allcott, H., Gentzkow, M.: Social media and fake news in the 2016 election. Journal of Economic Perspectives 31, 211–236 (2017)
2. Reuter, C., Hartwig, K., Kirchner, J., Schlegel, N.: Fake News Perception in Germany: A Representative Study of People's Attitudes and Approaches to Counteract Disinformation. 14th International Conference on Wirtschaftsinformatik (WI 2019) (forthcoming)
3. Forbes: Can 'Fake News' Impact the Stock Market?, https://www.forbes.com/sites/kenrapoza/2017/02/26/can-fake-news-impact-the-stock-market (Accessed: 5.05.2018)
4. Reuter, C.: Information Technology for Peace and Security - IT-Applications and Infrastructures in Conflicts, Crises, War, and Peace. Springer Verlag, Wiesbaden (Germany) (2019)
5. Asadullah, A., Kankanhalli, A., Faik, I.: Understanding Users' Intention to Verify Content on Social Media Platforms. 22nd Pacific Asia Conference on Information Systems, 251 (2018)
6. Müller, P., Denner, N.: Was tun gegen „Fake News"? (in German). Report. Friedrich-Naumann-Stiftung für die Freiheit. (2017)
7. Stanoevska-Slabeva, K.: Teaching Social Media Literacy with Storytelling and Social Media Curation. 23rd Americas Conference on Information Systems (2017)
8. Aufderheide, P.: Media Literacy. A Report of the National Leadership Conference on Media Literacy. ERIC (1993)
9. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design science in information systems research. Management Information Systems Quarterly 28, 725–730 (2008)
10. Viviani, M., Pasi, G.: Credibility in social media: opinions, news, and health information—a survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 7, e1209 (2017)
11. Conroy, N.J., Rubin, V.L., Chen, Y.: Automatic deception detection: Methods for finding fake news. In: Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community, p. 82 (2015)
12. Gupta, A., Kumaraguru, P., Castillo, C., Meier, P.: TweetCred: Real-Time Credibility Assessment of Content on Twitter. 6th International Conference on Social Informatics, 228–243 (2014)
13. Rehm, G.: An Infrastructure for Empowering Internet Users to handle Fake News and other Online Media Phenomena. In: International Conference of the German Society for Computational Linguistics and Language Technology, pp. 216–231 (2017)
14. Pariser, E.: The filter bubble: How the new personalized web is changing what we read and how we think. Penguin (2011)
15. Kim, A., Dennis, A.R.: Says Who?: How News Presentation Format Influences Perceived Believability and the Engagement Level of Social Media Users. 51st Hawaii International Conference on System Sciences, 497 (2017)

16. Nickerson, R.S.: Confirmation bias: A ubiquitous phenomenon in many guises. Review of general psychology 2, 175 (1998)
17. Kahne, J., Bowyer, B.: Educating for democracy in a partisan age: Confronting the challenges of motivated reasoning and misinformation. American Educational Research Journal 54, 3–34 (2017)
18. Mihailidis, P., Viotty, S.: Spreadable spectacle in digital culture: Civic expression, fake news, and the role of media literacies in "post-fact" society. American Behavioral Scientist 61, 441–454 (2017)
19. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: Proceedings of the 20th international conference on World wide web, pp. 675–684 (2011)
20. Stieglitz, S., Bruns, A., Krüger, N.: Enterprise-related crisis communication on Twitter. 12th International Conference on Wirtschaftsinformatik (WI 2015), 917–932 (2015)
21. Baumann, A., Krasnova, H., Veltri, N.F., Ye, Y.: Men, Women, Microblogging: Where Do We Stand? 12th International Conference on Wirtschaftsinformatik (WI 2015), pp. 857–871 (2015)
22. Morris, M.R., Counts, S., Roseway, A., Hoff, A., Schwarz, J.: Tweeting is believing?: understanding microblog credibility perceptions. In: Proceedings of the ACM 2012 conference on computer supported cooperative work, pp. 441–450 (2012)
23. Wanas, N., El-Saban, M., Ashour, H., Ammar, W.: Automatic scoring of online discussion posts. In: Proceedings of the 2nd ACM workshop on Information credibility on the web, pp. 19–26 (2008)
24. Weimer, M., Gurevych, I., Mühlhäuser, M.: Automatically assessing the post quality in online discussions on software. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pp. 125–128 (2007)
25. Weerkamp, W., Rijke, M.: Credibility improves topical blog post retrieval. Proceedings of ACL-08: HLT, 923–931 (2008)
26. van Someren, M.W., Barnard, Y.F., Sandberg, J.A.C.: The think aloud method: a practical approach to modelling cognitive. Academic Press, London (1994)
27. Sunshine, J., Egelman, S., Almuhimedi, H., Atri, N., Cranor, L.F.: Crying Wolf: An Empirical Study of SSL Warning Effectiveness. 18th USENIX security symposium, pp. 399–416 (2009)