Distinguishing Between Truth and Fake

Using Explainable AI to Understand and Combat Online Disinformation

Isabel Bezzaoui, Jonas Fegert, Christof Weinhardt
Information Process Engineering
FZI Forschungszentrum Informatik
Karlsruhe/Berlin, Germany
e-mail: bezzaoui@fzi.de, fegert@fzi.de, weinhardt@fzi.de

Abstract — **Disinformation campaigns have become a major threat to democracy and social cohesion. Phenomena like conspiracy theories promote political polarization; they can influence elections and lead people to (self-)damaging or even terrorist behavior. Since social media users and even larger platform operators are currently unready to clearly identify disinformation, new techniques for detecting online disinformation are urgently needed. In this paper, we present DeFaktS, an Information Systems research project, which takes a comprehensive approach to both researching and combating online disinformation. The project develops a data pipeline in which (i) messages are extracted in large quantities from suspicious social media groups and messenger groups with the help of annotators. Based on this corpus, a Machine Learning-based System (ii) is trained that can recognize factors and stylistic devices characteristic of disinformation, which will be used for (iii) an explainable artificial intelligence that informs users in a simple and comprehensible way about the occurrence of disinformation. Furthermore, in this paper an interdisciplinary multi-level research approach focusing on media literacy and trust in explainable artificial intelligence is suggested in order to operationalize research on combating disinformation.**

Keywords - Fake news; disinformation detection; machine learning-based systems; design science research.

I. INTRODUCTION

As the major news source of today, social media channels and online news portals suffer from non-fact-based reporting and opinion dissemination [1]. Spreading virally, disinformation poses a central threat to the political process and social cohesion. Disinformation is defined as false information, spread with the intention to deceive. Fake news is an example of disinformation, which is why we use these two terms interchangeably [2]. It influences elections, and tempts people to engage in (self-)damaging or even terrorist behavior. Accordingly, it displays a generally undesirable phenomenon in public information and opinion-forming processes [3][4]. Besides political radicalization [5], vaccination boycotts are increasingly attributed to disinformation campaigns [6][7]. Therefore, on the one hand, there is a need for a comprehensive understanding of their mechanisms and spread, and on the other hand, based on this, methods to combat them. People are naturally inclined to consume content with which they are familiar (familiarity bias), whose authors are similar to them (similarity bias), or whose statements they agree with (confirmation bias). In particular, confirmation bias is a decisive factor in the spread of disinformation [8]. Platforms, such as WhatsApp and Telegram in particular play a major role here and could take many preventive measures. They generally lack the appropriate approaches for this, because more emotional arousal and dissent lead to more activity on the platform, and in turn generate more advertising revenue [9][10]. Even though Twitter, for example, is experimenting with fact checks, these are far from sufficient to limit the spread of fake news as they do not operate across platforms. Therefore, DeFaktS intends to empower actual users across various platforms to critically question news and social media posts. For this purpose, the project will develop an artefact for a participation platform that aims to combat online disinformation campaigns and foster critical media literacy among users by informing them about the occurrence of fake news in a transparent and trustworthy way. The paper is structured as follows: Section II will give an introductory overview on the current knowledge base on the combat of disinformation as well as the concepts of critical media literacy and trust. Subsequently, the scientific method and first research activities in the project will be presented in Section III. Finally, the paper concludes with a summary of the project's research endeavors and an outlook on future work related to the project in Section IV.

II. THEORETICAL FOUNDATION

A. Combating Disinformation Using Machine Learning-Based Systems

The fact that nowadays almost anyone can publish content on the internet not only increases the possibility of social participation - it also creates new opportunities for

spreading disinformation and propaganda. The COVID-19 pandemic has already produced a flood of false reports and demonstrated the importance of being able to distinguish reliable information from half-truths and fake news, and most recently the war in Ukraine also demands a special confrontation with fake news [11]. Currently, research on fake news detection using Machine Learning-based Systems (MLS) is a rapidly expanding field that spans numerous disciplines, including computer science, social science, psychology, and information systems [12]-[14]. Synoptically, empirical efforts to detect and combat disinformation can be divided into four categories: data-oriented, feature-oriented, model-oriented and application-oriented [1]. The majority of methods concentrate on extracting multiple features, putting them into classification models, such as naive Bayes, logistic regression, or decision trees, and then selecting the best classifier based on performance [15]-[18]. What is missing from the previous work, however, are empirical evaluations of when the classifiers are put into practice with real users and of what benefits and impact the presented tools may have. For instance, Guess et al. [19] showed that promoting media literacy can help people judge the accuracy of online content more accurately. Their findings suggest that a lack of critical media literacy is a major factor in why people fall prey to disinformation. Pennycook and Rand [20] found that susceptibility to fake news is driven mostly by poor critical thinking rather than by partisan bias per se. Thus, in order to counter false news, more critical media competence is needed on the part of users. From this point of view, it seems crucial to investigate the potential of MLS detection tools for promoting critical media literacy among social media users.

Furthermore, previous research has demonstrated the importance of trust for the acceptance and perceived usefulness of ICT tools, and MLS in particular [21][22]. Trust is one of the vital components to fostering active, engaged and informed citizens [23]. Transparency is therefore an important aspect when it comes to dealing with disinformation. In this regard, the challenge of how to positively affect trust when developing tools for fake news detection arises. The implementation of an XAI-approach into the development process seeks to make the system's internal dynamics more transparent, as well as the analysis' conclusions more understandable and hence trustworthy to the user. These observations give rise to the need to examine the effect of XAI (Explainable Artificial Intelligence) elements on user trust and thus acceptance and perceived usefulness of the final tool. In order to fill the two above-mentioned research gaps, we would therefore like to address the following research questions in the DeFaktS project:

How to design an artifact for the detection of online disinformation that helps to foster an informed and critically thinking citizenry?

i. (How) Does the tool promote critical media literacy by helping users identify disinformation more accurately?
ii. (How) Does the tool's XAI-component help users trust the algorithm's assessment?

B. Critical Media Literacy

Disinformation is producing uncertainty in the fact-checking process, endangering the public's ability to make informed decisions [24]. In order to foster a critical comprehension of both manipulative communications and the internet as a distribution medium, users must have broad knowledge and a deeper understanding of social media functionalities [25]. Critical media literacy encourages people to consider why a message was sent and where it came from [26]. Following Kellner and Share [27], critical media literacy entails developing skills in analyzing media codes and conventions, and the ability to critique stereotypes, dominant values, and ideologies, as well as the competence to interpret media texts' multiple meanings and messages. Furthermore, it assists individuals to use media responsibly, to discern and assess media content, to critically examine media forms, to explore media effects, and to deconstruct alternative media. However, systematic evaluation of positive or potential non-intended negative effects of the usage of MLS fake news detection tools on the cultivation of critical media literacy is scarce [28]. Schmitt et al. [28] define three dimensions of critical media literacy that can be referred to the critical handling of online disinformation:

i. Awareness: "Awareness" in this case means awareness of the existence of disinformation. This includes knowledge of various forms of disinformation (disinformation in picture, text, or video form, distorted articles, and political pseudo-press) as well as a deeper understanding of how media, and online media in particular, operate.
ii. Reflection: Reflection in the context of critical media literacy is about applying analytical criteria to internet content and determining whether or not it is deceptive. The conscious consideration ("reflection") of content with the character of news is relevant, the thorough thinking before an article is liked, shared or the claim of a headline is taken at face value. As a result, reflection utilizes an individual's knowledge, abilities, and attitudes to critically evaluate (media-communicated) information based on specific criteria including credibility, source, and quality.
iii. Empowerment: Individuals' confidence in their ability to detect manipulative messages, participate in social discourses, and actively position themselves against disinformation is cultivated

through empowerment strategies and methods. In this context, empowerment can be defined as a certain form of behavior that encompasses a person's ability to recognize and express doubts about specific content as well as express their own thoughts.

In the DeFaktS project, these three dimensions will be used to investigate whether and to what extent the developed MLS can make a positive contribution to the cultivation of critical media competence among social media users. To this end, it will be analyzed whether and to what extent awareness, reflection, and empowerment are strengthened through the use of the artifact.

C. *Trust*

Niklas Luhmann [29] understands trust in the broadest sense as an elementary component of social life, interpreting it as a form of security, which can only be gained and maintained in the present. First and foremost, trust is needed to reduce a future of more or less undetermined complexity. According to Luhmann's understanding, the constant technical progress of society brings with it a simultaneous increase in complexity, which subsequently results in an increased need for trust. Thus, trust is a necessary condition to live and act with growing complexity in relation to modern events and dynamics [29]. However, trust is severely shaken by negative experiences [30], for instance experienced deception through disinformation. As MLS systems and algorithms become more complex, people increasingly regard them as "black boxes" that defy comprehension in the sense that understanding an MLS's decision requires growing amounts of specialized expertise and knowledge. Non-expert end-users are not able to retrace how the algorithmic code cascades led to a given decision [31]. Accordingly, there has been increased demand to offer the proper explanation for how and why a particular result was obtained [32]. Recent empirical evidence on algorithm acceptance [33] insinuates that explainability plays a heuristic role in algorithm and MLS service acceptance. Currently, however, research gives light to a controversy over whether the implementation of XAI-features actually helps increase user-trust or not. Shin [34] analyzed the impact of explainability in MLS on user trust and attitudes towards MLS and concluded that the inclusion of causability and explanatory features in MLS assists to increase trust as it helps users understand the decision-making process of MLS algorithms by providing transparency and accountability. In contrast, through their experiment on transparency and trust in MLS, Schmidt et al. [35] found that transparency features can actually affect trust negatively. These recent contradictory observations give rise to the need for further investigation of the effect of explainability on user trust. In the DeFaktS project, this research gap will be addressed through the evaluation of whether, and if so which, XAI elements increase user trust in the application.

III. Method and First Activities in Design Science Research

The goal of DeFaktS is to develop an artifact that is as close as possible to the needs of the subsequent user so that it contributes precisely to solving the above-mentioned issues. To implement this, the project is embedded in a Design Science Research (DSR) approach according to Peffers et al. [36]. DSR provides an adequate framework for contributing to both the theory and practice of solving real-world problems as it helps generate prescriptive knowledge on how to effectively design and deploy novel solutions to relevant problems [37]. The chosen approach divides the research process into six steps: problem identification and motivation, definition of the objectives for a solution, design and development, demonstration, evaluation, and communication [36]. Based on reviewing relevant literature from various disciplines, such as computer science, social science and information systems, we identified the problem and formulated the motivation to contribute to a solution (1). Inferred from the problem specification, the objectives of a solution with an emphasis on fostering critical media literacy and user trust were defined (2). On the basis of Step One and Two, the design science artifact, an XAI-tool detecting and warning social media users about online disinformation, will be created (3). For this purpose, trained annotators will extract and label messages from suspicious social media groups in large quantities. Subsequently, the data corpus will be provided for the training of a Machine Learning-based System to detect factors and stylistic devices characteristic of fake news. Finally, this system will be used for the XAI component that informs users in a simple and transparent way about the occurrence of these factors. In the fourth step, conducting a field study, the performance of the DeFaktS artifact will be demonstrated in a real world scenario (4). In this way, we will test whether our artifact serves to solve the identified problem in a suitable context. By conducting additional experimental studies in a lab environment, the efficiency and effectiveness of the tool will be evaluated. This step will help to assess whether the artifact factually helps to promote critical media literacy and increases user trust. Depending on the empirical results, possible iterations in the design and development process will follow (5). Finally, we will communicate our findings from Step 1 to 5 in scholarly and professional publications as well as at conferences and other suitable events. Furthermore, we will enable companies to create corresponding products for customers from business and civil society through an API, ensuring the artifact's sustainability (6).

Currently, researchers are concerned with Step Two and Three: The development of a 'Fake News Taxonomy' that

entails linguistic features and dimensions of disinformation content shall facilitate and ensure the quality of the data labeling process. One of the difficulties in detecting false news is that some terms and expressions are unique to a particular type of event or topic. When a fake news classifier is trained on fake versus real articles based on a certain event or topic, the classifier learns event-specific features and may not perform well when used to identify fake versus real articles based on a different type of event. As a result, fake news classifiers must be generalized to be event-independent [2]. Another challenge is that the majority of datasets are in English, and German-language datasets are scarce [38]. These observations call for the creation of a taxonomy of fake news that encompasses broad and event-independent dimensions and characteristics of disinformation, which is still specific enough to precisely identify and label deceiving content. The final taxonomy will display a design artifact in and for itself that will be demonstrated and evaluated within the labeling process. After some possible iterations, the artifact will be made accessible to other researchers through scientific publications or open access services.

IV. Conclusion and Future Work

In this research-in-progress, we contribute to the knowledge base of fake news detection using MLS by developing an XAI-artifact and evaluating its performance in the context of fostering critical media literacy and trust among social media users. The innovative aspects of DeFaktS are multifold: Non-expert users shall be enabled to understand, trust, and utilize the tool's interpretation and explanation of detection results. Further, the DeFaktS-tool shall increase overall critical thinking and awareness of online disinformation, cultivating an informed citizenry and fostering political participation. The presented project is to be understood as work in progress during which the six steps of design science research are followed and critically evaluated simultaneously. For now, this paper intends to show the scientific community initial approaches to researching and combating online disinformation campaigns using Machine Learning-based Systems while remaining receptive to future developments in empiricism and civil society.

References

[1] K. Shu, A. Sliva, S. Wang, J. Tang, L. Huan, "Fake news detection on social media: A data mining perspective", *ACM SIGKDD Explorations Newsletter*, 19, pp. 22-36, 2017.

[2] K. Shu, A. Bhattacharjee, F. Alatawi, T. H. Nazer, K. Ding, M. Karami, H. Liu, "Combating disinformation in a social media age", *WIREs Data Mining and Knowledge Discovery*, *10*, pp. 1–23, 2020.

[3] D. McQuail, "Media performance: Mass communication and the public interest", *Thousand Oaks*, CA: Sage. M. Young, The Technical Writer's Handbook, Mill Valley, CA: University Science, 1992.

[4] J. Strömbäck, "In search of a standard: Four models of democracy and their normative implications for journalism", *Journalism Studies, (6:3)*, pp. 331-345, 2005.

[5] J. Groshek, and K. Koc-Michalska, "Helping populism win? Social media use, filter bubbles, and support for populist presidential candidates in the 2016 US election campaign", *Information Communication and Society, (20:9)*, pp. 1389-1407, 2017.

[6] H. Holone, "The filter bubble and its effect on online personal health information", *Croatian Medical Journal, (57:3*), pp. 298–301, 2016.

[7] K. Sharma, Y. Zhang, Y. Liu, "COVID-19 vaccines: characterizing misinformation campaigns and vaccine hesitancy on twitter", Retrieved May 2022. *arXiv preprint arXiv:2106.08423*, 2021.

[8] E.C. Tandoc Jr, "The facts of fake news: A research review", *Sociology Compass*, *13*(9), e12724, pp. 1-9, 2019.

[9] L. Munn, "Angry by design: toxic communication and technical architectures", *Humanities and Social Sciences Communications*, *7*(1), pp. 1-11, 2020.

[10] K. Nelson-Field, E. Riebe, K. Newstead, "The emotions that drive viral video", *Australasian Marketing Journal*, *21*(4), pp. 205–211, 2013.

[11] J. Delcker, Z. Wanat, M. Scott, „The coronavirus fake news pandemic sweeping WhatsApp", *Politico,* Retrieved May 2022 from https://www.politico.eu/article/the-coronavirus-covid19-fake-news-pandemic-sweeping-whatsapp-misinformation/, 2020.

[12] S. Yu, and D. Lo, "Disinformation detection using passive aggressive algorithms", *ACM Southeast Conference, Session 4*, p. 324f, 2020.

[13] P. K. Verma, P. Agrawal, I. Amorim, R. Prodan, "WELFake: Word embedding over linguistic features for fake news detection", *IEEE Transactions on Computational Social Systems, 8(4),* pp. 881–893, 2021.

[14] M. Mahyoob, J. Al-Garaady, M. Alrahaili, "Linguistic-based detection of fake news in social media." Forthcoming, *International Journal of English Linguistics, 11(1),* pp. 99-109*,* 2020.

[15] H. Alsaidi, and W. Etaiwi, "Empirical evaluation of machine learning classification algorithms for detecting COVID-19 fake news", *Int. J. Advance Soft Compu. Appl, 14(1),* pp. 49-59, 2022.

[16] W. H. Bangyal et al., "Detection of Fake News Text Classification on COVID-19 Using Deep Learning Approaches", *Computational and Mathematical Methods in Medicine,* pp. 1-13, 2021.

[17] L. Bozarth, and C. Budak, "Toward a better performance evaluation framework for fake news classification", *Proceedings of the international AAAI conference on web and social media,* 14, pp. 60–71, 2020.

[18] C. Lai et al., "Fake news classification based on content level features", *Applied Sciences, 12(3)*, p. 1116, 2022.

[19] A. M. Guess et al., "A digital media literacy intervention increases discernment between mainsitream and false news in the United States and India", *PNAS*, *117*(27), pp. 15536–15545, 2022.

[20] G. Pennycook, and D. G. Rand, "Lazy, not biased: Suceptibility to partisan news is better explained by lack of reasinong than by motivated reasoning", *Cognition*, pp. 1–12, 2018.

[21] D. Ribes Lemay et al., "Trust indicators and explainable AI: A study on user perceptions", *IFIP Conference on Human-Computer Interaction - INTERACT 2021*, pp. 662–671, 2021.

[22] K. Siau, and W. Wang, "Building trust in artificial intelligence, machine learning, and robotics", *Cutter Business Journal*, *31*(2), pp. 47-53, 2018.

[23] P. Dahlgren, "Media and political engagement: Citizens, communication, and democracy", Cambridge: Cambridge University Press, 2009.

[24] S. M. Jang, and J. K. Kim, "Third person effects of fake news: Fake news regulation and media literacy interventions", *Computers in Human Behavior*, *80*, pp. 295–302, 2018.

[25] D. Rieger et al., "Propaganda und Alternativen im Internet - Medienpädagogische Implikationen. Propaganda and Alternatives on the Internet - Media Pedagogical Implications", *merz / medien + erziehung,* (3), pp. 27-35, 2017.

[26] D. Kellner, and J. Share "Critical media literacy, democracy, and the reconstruction of education", In D. Macedo & S. R. Steinberg, *Media Literacy: A Reader*, Peter Lang Publishing, pp. 3-23, 2007.

[27] D. Kellner, and J. Share, "Toward critical media literacy: Core concepts, debates, organizations, and policy", *Discourse: studies in the cultural politics of education*, *26*(3), pp. 369–386, 2005.

[28] J. B. Schmitt, D. Rieger, J. Ernst, H. J. Roth, „Critical media literacy and Islamist online propaganda: The feasibility, applicability and impact of three learning arrangements", *International Journal of Conflict and Violence*, *12*, pp. 1–19, 2018.

[29] N. Luhmann, „Vertrauen", Trust (5), UVK, 2014.

[30] F. Schwerter, and F. Zimmermann, „Determinants of trust: The role of personal experiences", *Games and Economic Behavior*, *122*, pp. 413-425, 2020.

[31] D. Castelvecchi, "Can we open the black box of AI?", *Nature*, *538*, pp. 20–23, 2016.

[32] M. Ter Hoeve et al., „Do news consumers want explanations for personalized news rankings?" *FARTEC*, pp. 1–6, 2017.

[33] D. Shin, B. Zhong, F. A. Biocca, "Beyond user experience: What constitutes algorithmic experiences?" *International Journal of Information Management*, *52*, pp. 1–11, 2020.

[34] D. Shin, "The effects of explainability and causability on perception, trust and acceptance: Implications for explainable AI", *International Journal of Human-Computer Studies*, *146,* pp. 1-11, 2021.

[35] T. Schmidt, F. Biessmann, T. Teubner, „Transparency and trust in artificial intelligence systems", *Journal of Decision Systems*, *29*(4), pp. 260–278, 2020.

[36] K. Peffers, T. Tuunanen, M. A. Rothenbergre, S. Chatterjee, "A design science research methodology for information systems research", *Journal of Management Information Systems*, *24*(3), pp. 45–77, 2007.

[37] J. Vom Brocke, A. Hevner, A. Maedche, „Introduction to design science research", In J. Vom Brocke, A. Hevner, & A. Maedche (edt.), *Design Science Research*. Cases, Cham, pp. 1-18, 2020.

[38] D. Schreiber, C. Picus, D. Fischinger, M. Boyer, „The defalsif-AI project: Protecting critical infrastructures against disinformation and fake news/Das Projekt defalsif-AI: Schutz kritischer Infrastrukturen vor Desinformation und Fake News." *Elektrotechnik und Informationstechnik*, *Vol. 138 (7)*, pp. 480–484, 2021.