IEAI Technical University of Munich
Munich Center for Technology in Society
Institute for Ethics in Artificial Intelligence

Research Brief – December 2020

Negative Dynamics on Social Media and their Ethical Challenges for AI

by Wienke Strathern and Jürgen Pfeffer

Social media are platforms on which millions of people gather information, interact and form opinions. More recently, we have observed a surge in negative opinion formation. Online firestorms, fake news and hate speech have shaken our beliefs and hopes about the positive power of social media to their very foundations. While negative emotions are in the core of human behavior, algorithms on social media, enhanced by Artificial Intelligence (AI) can produce and reinforce new dynamics. Enhancing fairness, minimizing biases and, consequently, reducing the negative dynamics on social media requires an understanding of the driving forces of these dynamics.

1. What are Online Firestorms?

As social media platforms with hundreds of millions of users interacting in real-time on topics and events all over the world, social media networks are social sensors for online discussions and are known for quick and often emotional disputes. Online firestorms can be defined as the sudden discharge of large quantities of messages containing negative word-of-mouth and complaint behavior against a person, company or group in social media networks (Pfeffer et al. 2013). In social media networks, they are seemingly arbitrarily occurring outrages towards people, companies, media campaigns and politicians. But moral outrages can create an excessive collective aggressiveness against one single argument, one single word or one action of a person, resulting in hateful speech. The negative dynamics often start with a collective "against the others" (Strathern et al. 2020).

In social media networks, online firestorms are seemingly arbitrarily occurring outrages towards people, companies, media campaigns and politicians.

In social media, negative opinions about products or companies are formed by and propagated via thousands or millions of people within hours. Furthermore, massive negative online dynamics are not only limited to the business domain, but they also affect organizations and individuals in politics. Even though online firestorms are a new phenomenon, their dynamics are similar to the way in which rumors are circulated. In 1947, Gordon Allport and Leo Postman defined a rumor as a “proposition for belief, passed along from person to person, usually by word-of-mouth, without secure standards of evidence being presented” (Allport 1947).

2. The Problem with AI and Negative Word-of-Mouth Dynamics

When people are active on social media, they act in a socio-technical system that is mediated and driven by AI-powered algorithms. The goal of social media platforms is to keep users engaged and to maximize their time spent on the platform. Highly engaged users who spend a lot of time on platforms are the core of a social media business model that is based on selling more and better targeted ads. But the question is always: which content will be interesting for a particular user? To answer this, recommender systems are developed to increase the chance that a user will click on a suggested link and read its content. These recommender algorithms incorporate socio-demographic information, but also data of a user’s previous activity. Furthermore, behavioral data of alters (friends) of a user are also used to suggest new content. Social scientists have studied the driving forces of social relationships for decades, i.e. why do people connect with each other. **Homophily** and **transitivity** are the most important factors for network formation. Homophily means that your friends are similar to yourself. They like similar things and are interested in similar topics (McPherson et al 2001). Transitivity describes the fact that a person’s friends are often connected among each other. Combining these two aspects results in the fact that most people are embedded in personal networks with people that are similar to themselves and who are to a high degree connected among each other.

Unfortunately, the above-described forces of how humans create networks combined with AI-driven recommender systems have problematic implications. Recommender systems filter the content that is presented on social media and suggest new “friends” to us. As a result, filter bubbles (Pariser 2011) are formed around individuals on social media, i.e. they are connected to like-minded people and familiar content. The lack of diversity in access to people and content can easily lead to polarization. If we now add another key characteristic of social media, abbreviated communication with little space for elaborate exchange, a perfect breeding ground for online firestorms emerges. Imagine a couple of people disliking a statement or action of a politician, celebrity or any private individual and these people voicing their dislike aggressively on social media. Their online peers, who most likely have similar views (see above), will easily and quickly agree by sharing or re-tweeting the discontent. Within hours, these negative dynamics

can reach tens of thousands of users. Voilá, an online firestorm is born.

3. Potential role for Artificial Intelligence to tackle Online Firestorms

In our projects[1], we apply methods from social network analysis and artificial intelligence/ machine learning to better understand the driving factors of social media group level phenomena like online firestorms that lead to negative dynamics. Our goal is to develop approaches for how to detect, react to and possibly mitigate these dynamics early on. In our paper “Against the Others! Detecting Moral Outrage in Social Media Networks", we explored the outbreak of online firestorms on Twitter. We used data from Twitter to investigate the starting points of several firestorm outbreaks. The paper aims to determine whether it is possible to detect the outbreak of a firestorm. Is there a changing point? How can the features that cause moral outrage be distinguished? In order to examine such challenges, we have developed a method to detect the point of change that systematically identifies linguistic cues contained in the tweets. This work is fundamental for detecting the pattern of the spread of negative dynamics and could help individuals, companies and governments to mitigate hate in social media networks.

Traditional detection of outrage (e.g. hate speech) is based on identification of pre-defined keywords, while the context in which certain topics and words are being used has to be almost disregarded. To name just one extreme example, hate groups have managed to escape keyword-based machine detection through clever combinations of words, misspellings, satire and coded language (Udupa 2020). The focus of our analysis is on more complex lexical characteristics in texts, which we apply as a basis for automated predictions and possible early interventions using AI-enabled tools.

We tackled the question of anomaly detection in a network by exploring major features that indicate the outbreak of a firestorm, hence, the starting point of possible collective changing behavior. On this account, examining the change of vocabulary to explore firestorm data revealed how people connect with each other to form an outrage. Interestingly, a change of personal pronouns is a very good indicator for online firestorms. On social media, people love to talk about themselves. But when they turn their outrage against somebody, the occurrence of ‘I, me, ...’ decreases significantly, and at the same time the mentioning of a user who is the target of a firestorm increases strongly. The individual user moves away from him- or herself as a subject, the perspective changes.

4. Ethical Considerations using AI Methods to Tackle Online Firestorms

Our research touches a variety of ethical topics (see final thoughts). Besides that, our own research faces ethical challenges every step of the way. Many researchers rely on social media data as a resource. A current topic of debate within the computational social science research community involves the ethical (or even legal) implications of collecting data in ways that violate Terms of Service (TOS) (Fiesler et al. 2020), privacy, personal data rights or free speech (Sunstein 2017, Joergensen 2019).

a. Privacy

The volume and relativity of data collection will keep privacy at the forefront as one of the most significant legal issues that AI users will face going forward. AI systems use vast amounts of data. Therefore, as more data is used, more questions are raised. Who owns the data shared between AI developers and users? Can data be shared for research purposes or even sold? Should this shared data be de-identified to protect privacy concerns? Is the intended use of data

[1] This includes the IEAI project - Online Firestorms and Resentment Propagation on Social Media: Dynamics, Predictability, and Mitigation

appropriately disclosed and compliant with legislation?

b. Personal Data Rights

The prediction of possible opinions and dynamics further touches the question of personal data rights and how to deal with them. Can we really assume that people willingly share their opinions in social media and that they are aware of its consequences?

c. Free speech

When it comes to identifying statements at an early stage and stopping them if necessary, we are very much in the realm of free speech. Since online firestorms happen globally, across national borders, the question arises as to the ethical and legal basis of this action.

5. Final Thoughts

The goals of our research will provide insights of ethical relevance by discussing responsibility, delegation and control mechanisms in human-AI interacting systems. Understanding driving forces of opinion dynamics will enhance **fairness, minimizing biases** and consequently, can help to reduce the negative dynamics on social media. Unmasking mal-actors that try to intentionally create negative dynamics on social media by conning, tricking and manipulating the crowd, will **enhance transparency, explicability and use of data** and furnish fundamental understanding to public regulators for the **governance, regulation and sustainability** of proper and improper use of social media.

*Wienke Strathern and Prof. Dr. Jürgen Pfeffer conduct research at the Professorship for Computational Social Science and Big Data at the TUM School of Governance. They work on the current IEAI research projects: *Online Firestorms and Resentment Propagation on Social Media: Dynamics, Predictability* and *Mitigation* and *Online-Offline Spillovers – Potential real-world implications of online manipulation*.

6. References

G.W. Allport, and L. Postman. The psychology of rumor. New York: Henry Holt and Company, 1947.

C. Fiesler, N. Beard, & B.C. Keegan. No Robots, Spiders, or Scrapers. Legal and Ethical Regulation of Data Collection Methods in Social Media Terms of Service. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*(1), 187-196, 2020.

R. F. Joergensen (Editor): Human Rights in the Age of Platforms. The MIT Press 2019.

D. Lazer, and A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne. Computational Social Science. *Science* 323, 5915, pages 721-723, 2009.

D. Lazer, and A. Pentland, D.J. Watts, S. Aral, S. Athey, N. Contractor, D. Freelon, S. Gonzalez-Bailon, G. King, Gary, H. Margetts, A. Nelson, M. Salganik, M. Strohmaier, A. Vespignani, and C. Wagner, Claudia. Computational social science: Obstacles and opportunities. *Science* 369, 6507, pages 1060-1062, 2020.

M. McPherson, L. Smith-Lovin, and J.M. Cook: Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27:415-444, 2001.

E. Pariser. The Filter Bubble: What the Internet Is Hiding from You. Penguin Press, New York, 2011.

J. Pfeffer, T. Zorbach, and K. M. Carley. Understanding online firestorms: Negative word-of-mouth dynamics in social media networks. *Journal of Marketing Communications*, pages 1–12, 2013.

C.R. Sunstein. #Republic: Divided Democracy in the Age of Social Media. Princeton University Press, 2017.

S. Udupa: Artificial Intelligence and the Cultural Problem of Extreme Speech. *Social Science Research Council*, (20 December 2020), online: https://items.ssrc.org/disinformation-democracy-and-conflict-prevention/artificial-intelligence-and-the-cultural-problem-of-online-extreme-speech/