This article has been accepted for publication in 29th ACM Conference on User Modeling, Adaptation and Personalization. Citation information: https://doi.org/10.1145/3450614.3463353

Towards Continuous Automatic Audits of Social Media Adaptive Behavior and its Role in Misinformation Spreading

JAKUB SIMKO, Kempelen Institute of Intelligent Technologies, Slovakia
MATUS TOMLEIN, Kempelen Institute of Intelligent Technologies, Slovakia
BRANISLAV PECHER, Kempelen Institute of Intelligent Technologies, Slovakia
ROBERT MORO, Kempelen Institute of Intelligent Technologies, Slovakia
IVAN SRBA, Kempelen Institute of Intelligent Technologies, Slovakia
ELENA STEFANCOVA, Kempelen Institute of Intelligent Technologies, Slovakia
ANDREA HRCKOVA, Kempelen Institute of Intelligent Technologies, Slovakia
MICHAL KOMPAN[*], Kempelen Institute of Intelligent Technologies, Slovakia
JURAJ PODROUZEK, Kempelen Institute of Intelligent Technologies, Slovakia
MARIA BIELIKOVA[†], Kempelen Institute of Intelligent Technologies, Slovakia

In this paper, we argue for continuous and automatic auditing of social media adaptive behavior and outline its key characteristics and challenges. We are motivated by the spread of online misinformation, which has recently been fueled by opaque recommendations on social media platforms. Although many platforms have declared to take steps against the spread of misinformation, the effectiveness of such measures must be assessed independently. To this end, independent organizations and researchers carry out audits to quantitatively assess platform recommendation behavior and its effects (e.g., filter bubble creation tendencies). The audits are typically based on agents simulating the user behavior and collecting platform reactions (e.g., recommended items). The downside of such auditing is the cost related to the interpretation of collected data (here, some auditors are advancing automatic annotation). Furthermore, social media platforms are dynamic and ever-changing (algorithms change, concepts drift, new content appears). Therefore, audits need to be performed continuously. This further increases the need for automated data annotation. Regarding the data annotation, we argue for the application of weak supervision, semi-supervised learning, and human-in-the-loop techniques.

CCS Concepts: • **Information systems** → **Personalization**; • **Human-centered computing** → *Human computer interaction (HCI).*

Additional Key Words and Phrases: audits, social media, personalization, recommendations, misinformation, filter bubbles

ACM Reference Format:
Jakub Simko, Matus Tomlein, Branislav Pecher, Robert Moro, Ivan Srba, Elena Stefancova, Andrea Hrckova, Michal Kompan, Juraj Podrouzek, and Maria Bielikova. 2021. Towards Continuous Automatic Audits of Social Media Adaptive Behavior and its Role in Misinformation Spreading. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '21 Adjunct), June 21–25, 2021, Utrecht, Netherlands.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3450614.3463353

[*] Also with slovak.AI.
[†] Also with slovak.AI.

UMAP '21 Adjunct, June 21–25, 2021, Utrecht, Netherlands

ACM ISBN 978-1-4503-8367-7/21/06…$15.00
https://doi.org/10.1145/3450614.3463353

This article has been accepted for publication in 29th ACM Conference on User Modeling, Adaptation and Personalization. Citation information: https://doi.org/10.1145/3450614.3463353

1 PLATFORM ACCOUNTABILITY VIA ADAPTIVE BEHAVIOR AUDITS

The factors of today's online misinformation spreading are (perhaps) known but are certainly hard to quantify. The spreading of false information is influenced by many parties involved: from content consumers (users) and content creators to social media platforms. They interact in a complex environment and exchange influences with the offline world. Thus, it is hard to pinpoint systemic reasons and make parties accountable. Despite that, adaptive behavior of social media platforms has been repeatedly identified as a contributing factor (as recommendation algorithms often prefer attention-keeping content) [22]. Pushed by public outcry as well as early regulatory initiatives (e.g., EU's Code of practice on disinformation[1]), platforms and their representatives are pledging to implement preventive measures [9]. However, these pledges are hard to evaluate as platforms remain closed. Currently, the efficiency of platforms' countermeasures is primarily self-evaluated, which is prone to subjectivity and are difficult to validate.

Independent and quantitative *auditing* of the social media platform adaptive behavior may lead to more transparency. An *audit* is a black-box probing technique independent of the platform's cooperation, where user behavior is simulated by an agent over a live platform, adaptive behavior of which is traced and analyzed. For example, using a new user account on a video-oriented platform, one may first visit (watch) some seed videos (e.g., about some misinformation-linked topic) and then analyze the recommendations given by the platform (e.g., do they promote misinformation). If repeated over multiple times, topics, users, user characteristics, seed videos, or feedback options (views, likes, shares, etc.), a quantitative image about misinformation spreading or filter bubble creation on the platform can be acquired.

While existing audits are commendable, they suffer from several drawbacks. Most importantly, their results quickly become obsolete due to changes in: (1) platform content (new content is constantly being uploaded and is often preferred [4]), (2) adaptive algorithms [25], and (3) evolving policies (platforms implement specific mitigation features [1]). Methodological inconsistencies between various surveys are another disadvantage (e.g., differences in agent behavior or data annotation schemes), making the comparison between studies difficult. It is also hard to see trends, as studies are primarily conducted in short time periods.

We argue that auditing must be done in a more systematic fashion and continuously. Continuous auditing, i.e., permanent (or rather repeated) execution of probing activities over a live platform, would allow capturing trends occurring on the platform and prevent knowledge obsoletion. Moreover, we may even distinguish abrupt changes of platform behavior (caused likely by policy changes) from gradual changes (caused likely by learning of algorithms).

The scale of auditing implies a strong need for automation. The obvious role for automation is the user interaction with the platform, already done using software agents in many studies. However, more challenging (and needed) is to automate the data analysis, specifically the annotation of the observed content (e.g., Is this post conspiratorial? Is that video a pseudoscience?). In most cases, the annotation is performed by humans, limiting the scalability. Some automated techniques have already been tried in the literature [16, 19]. *In this paper, we argue for the application of weak supervision, semi-supervised learning, and human-in-the-loop techniques.* With it, we seek to tackle large volumes of data, lack of training labels, concept drifts, new topics, and a lack of agreement on labeling schemes.

Audits of social media can take various forms. Sandvig et al. [18] identified several options for data collection. In crowdsourcing audits, data is collected from real users in-the-wild. Social media applications are scraped as they are visited by users through the use of browser extensions and add-ons [19]. This provides an authentic picture of the users' experience. However, scaling the

[1] https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation

This article has been accepted for publication in 29th ACM Conference on User Modeling, Adaptation and Personalization.
Citation information: https://doi.org/10.1145/3450614.3463353

data collection to achieve a representative pool of users is challenging due to various ethical and legal issues, e.g. concerning privacy. Also, the uncontrolled nature of the environment makes data comparisons hard. To overcome this, recent works utilized bots to impersonate the behavior of users in a more controlled and uniform way in *sockpuppeting* audits [7, 16].

As an audit example, we can mention our ongoing project, which builds on existing research [7, 16]. It uses bots to investigate how misinformation filter bubbles are formed and how they may be bursted. YouTube is used as a case to analyze this behavior. To create misinformation filter bubbles, bots consume content (watch videos) belonging to a single stance within a misinformation topic (e.g., promoting the flat earth conspiracy). The seed data are hand-picked for each analyzed topic and stance. Bots impersonate real users by simulating specific web browsing behavior. This is defined by simple rules and mainly involves watching videos and navigating the interface. Explicit feedback such as adding likes to videos is used as well. Data is collected by scraping recommendations from visited websites and search results from predefined queries. The presence of misinformative content is annotated manually.

2 TOWARDS CONTINUOUS AUDITING

Continuous auditing is a well-known concept in the finance domain. Kogan et al. [12] defined continuous auditing as a type of auditing that produces results simultaneously with (or shortly after) the occurrence of relevant events. One of the motivating forces for continuous audits in finance has been the growth of fast online trading. The dynamic environment on social media brings up a similar need for continuous monitoring. Longitudinal algorithmic audits would allow comparing personalization along a different axis than audits capturing snapshots of single points in time [15].

Longitudinal audits have already been employed to audit search results [15] and news headlines from Google Top Stories [11]. However, longitudinal audits of personalization on social media platforms are largely missing. Relatively rare crowdsourcing audits with real users [19] and longitudinal studies [14] have posed challenges in retaining engaged participants and scaling outside of local communities. Sockpuppeting audits tend to be carried out as snapshots of single points in time [7, 16], but can also span over several months [20]. By simulation of a range of user behaviors and control of the conditions of the audit (geolocation, demographics, etc.), they are well-suited to the task.

Continuous audits bring several challenges. The platforms are dynamic, and some of their changes surface over a long time period. It is essential to distinguish between *endogenous* (changes in algorithms, policy decisions made by platforms) and *exogenous* factors (changes in content, external events, behavior of content creators) in audit results over time [15]. Related work proposed frameworks, e.g., to distinguish and quantify sources of bias in search results [13]. Similar tools will be needed for separating different sources of dynamism in audit results over time. Comparative studies across platforms and topics of audited content could also be helpful.

Furthermore, platforms frequently modify their user interfaces and services. Maintenance of agent scripting is necessary to keep up with the changes even during a short audit. To enable long term comparisons, continuous audits will need to focus on core features (that tend to be stable and common across platforms, e.g., endorsement option “like”). On the other hand, agents should be endowed with a *user-realistic behavior* to reflect real conditions and adaptation impacts. Some levels of fidelity can usually be achieved by conventional approaches (manually scripted scenarios [7, 16]). However, it could be supplemented by models learned on real user behavior (e.g. [10]).

This article has been accepted for publication in 29th ACM Conference on User Modeling, Adaptation and Personalization.
Citation information: https://doi.org/10.1145/3450614.3463353

3 TOWARDS AUTOMATIC ANNOTATION

For continuous auditing to be effective, it needs to be done with minimal human effort. This is directly opposed by the need to annotate large volumes of collected data. Despite some possible annotation re-use (as some of the content is featured multiple times on a platform), new content emerging everyday will make the already created annotations obsolete. Such prospects ask for automatic content annotation methods.

Only a few social media audits perform automatic annotation. Silva et al. [19] used CNN to automatically identify political ads on Facebook, achieving high accuracy on a balanced dataset. In Papadamou et al. [16], the annotation of YouTube videos was done using NN classifier trained on a small set of crowd-sourced human annotations (using features such as title, description, transcript or comments). Unfortunately, the accuracy was meager due to the presence of human bias in the annotations [16]. To counter this drawback, a multi-step approach to annotation could be used [3, 24]. The intermediate steps can be more rigorously defined, and thus, annotation subjectivity can be lowered. For example, a monolithic classification can be replaced by a two-step process. In a first step, the presence of a known misinformative claim (drawn from a catalog) would be determined. In the second step, the stance of the content piece towards that claim would be estimated [23]. Although using multi-step approaches increase the annotation requirements (multiple labels are needed per observation), these could be lowered by using the weak supervision paradigm [17]. *Weak labels* can often be automatically generated, which further eliminates human efforts and reduces the annotator biases.

The automatic annotation would also have to solve the concept drift problem: term and phrase meanings change and new ones emerge. Capturing the new meanings is crucial as it opens up opportunities for auditing preventive actions (such as those taken recently against COVID pandemic [1]). For the concept drift problem, we see human-in-the-loop approaches as the answer [2, 5, 6]. The concept drift detection can be done automatically by checking the confidence of the model or distribution and differences between observed samples [2, 5]. After detecting a concept drift, the optimal samples are selected and sent for human annotation to become part of training sets [2, 5, 6].

4 CONCLUSION AND DISCUSSION

Achieving proper accountability of social media platforms will require not only self-imposed rules or public regulations but also reliable, independent tools and methods for the platform behavior assessment. We envision one such family of tools oriented towards auditing of adaptive behavior, especially recommendations.

We argue that although recent auditing studies brought some insight into the behavior of large platforms, they become quickly outdated by content or policy changes. Therefore, the current practice of sporadic one-time studies has to shift into *continuous auditing*. With continuous auditing, we would be able to see trends in misinformative content spreading as well as effects of platform policy changes. Lot of data, generated by the audits, will need to be annotated through *automated means*, supplemented by human-in-the-loop solutions for hard-to-tackle cases, such as appearances of new content topics (e.g., COVID). The human annotators will also *curate the seed data pools* (e.g., search queries, videos, initial users/channels to follow, etc.) to representatively cover all relevant content domains and trending topics. Acquisition of these pools will also benefit from automated methods. The aim of automation is to minimize the human effort necessary, not their supervision, which is essential for the audits to be credible. But to avoid any substantial harms on human values and fundamental rights, proper ethical assessment should accompany every future research in this area [8].

The *credibility* of the independent audits is their key requirement. It goes hand in hand with their *reproducibility*. Thus, the methodology, seed data, collected data, and source codes used for auditing needs to be open (existing research platforms supporting open access to data, such as our platform MonAnt [21], may serve for this purpose). In addition, decisions of the automated methods need to be transparent and accompanied with explanations.

ACKNOWLEDGMENTS

This research was partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215.

REFERENCES

[1] Corey H Basch, Grace C Hillyer, Zoe C Meleo-Erwin, Christie Jaime, Jan Mohlman, and Charles E Basch. 2020. Preventive behaviors conveyed on YouTube to mitigate transmission of COVID-19: cross-sectional study. *JMIR public health and surveillance* 6, 2 (2020), e18807.

[2] Andrea Bontempelli, Fausto Giunchiglia, Andrea Passerini, and Stefano Teso. 2021. Human-in-the-loop Handling of Knowledge Drift. arXiv:2103.14874 [cs.LG]

[3] L. Borges, B. Martins, and P. Calado. 2019. Combining similarity features and deep representation learning for stance detection in the context of checking fake news. *Journal of Data and Information Quality (JDIQ)* 11, 3 (2019), 1–26.

[4] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep Neural Networks for YouTube Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems* (Boston, Massachusetts, USA) *(RecSys '16)*. ACM, New York, NY, USA, 191–198. https://doi.org/10.1145/2959100.2959190

[5] R. Elwell and R. Polikar. 2011. Incremental Learning of Concept Drift in Nonstationary Environments. *IEEE Transactions on Neural Networks* 22, 10 (2011), 1517–1531. https://doi.org/10.1109/TNN.2011.2160459

[6] Donald Honeycutt, Mahsan Nourani, and Eric Ragan. 2020. Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 63–72.

[7] Eslam Hussein, Prerna Juneja, and Tanushree Mitra. 2020. Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 048 (May 2020), 27 pages. https://doi.org/10.1145/3392854

[8] Insight. [n.d.]. The Assessment List for Trustworthy Artificial Intelligence. https://altai.insight-centre.org/. Accessed: 2021-04-12.

[9] Petros Iosifidis and Nicholas Nicoli. 2020. The battle to end fake news: A qualitative content analysis of Facebook announcements on how it combats disinformation. *International Communication Gazette* 82, 1 (2020), 60–81.

[10] Hamdi Kavak, Jose J. Padilla, Christopher J. Lynch, and Saikou Y. Diallo. 2018. Big Data, Agents, and Machine Learning: Towards a Data-Driven Agent-Based Modeling Approach. In *Proceedings of the Annual Simulation Symposium* (Baltimore, Maryland) *(ANSS '18)*. Society for Computer Simulation International, San Diego, CA, USA, Article 12, 12 pages.

[11] Anna Kawakami, Khonzodakhon Umarova, and Eni Mustafaraj. 2020. The Media Coverage of the 2020 US Presidential Election Candidates through the Lens of Google's Top Stories. *Proceedings of the International AAAI Conference on Web and Social Media* 14, 1 (May 2020), 868–877.

[12] Alexander Kogan, Ephraim F. Sudit, and Miklos A. Vasarhelyi. 1999. Continuous Online Auditing: A Program of Research. *Journal of Information Systems* 13, 2 (09 1999), 87–103. https://doi.org/10.2308/jis.1999.13.2.87

[13] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P. Gummadi, and Karrie Karahalios. 2017. Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media. In *Proc. of the 2017 ACM Conf. on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 417–432. https://doi.org/10.1145/2998181.2998321

[14] Stephen M. Mattingly et al. 2019. The Tesserae Project: Large-Scale, Longitudinal, *In Situ,* Multimodal Sensing of Information Workers. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. ACM, New York, NY, USA, 1–8.

[15] Danaë Metaxa, Joon Sung Park, James A. Landay, and Jeff Hancock. 2019. Search Media and Elections: A Longitudinal Investigation of Political Search Results. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 129 (Nov. 2019), 17 pages. https://doi.org/10.1145/3359231

[16] Kostantinos Papadamou, Savvas Zannettou, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Michael Sirivianos. 2021. " It is just a flu": Assessing the Effect of Watch History on YouTube's Pseudoscientific Video Recommendations. *arXiv preprint arXiv:2010.11638v3* (2021).

[17] Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and · Christopher Ré. 2020. Snorkel: rapid training data creation with weak supervision. *The VLDB Journal* 29 (2020), 709–730. https://doi.org/10.1007/s00778-019-00552-1

[18] Christian Sandvig, K. Hamilton, K. Karahalios, and C. Langbort. 2014. Auditing Algorithms : Research Methods for Detecting Discrimination on Internet Platforms. In *64th Annual Meeting of the International Communication Association.* Seattle, WA, 23 pages.

[19] Márcio Silva, Lucas Santos de Oliveira, Athanasios Andreou, Pedro Olmo Vaz de Melo, Oana Goga, and Fabricio Benevenuto. 2020. Facebook Ads Monitor: An Independent Auditing System for Political Ads on Facebook. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 224–234. https://doi.org/10.1145/3366423.3380109

[20] Larissa Spinelli and Mark Crovella. 2020. How YouTube Leads Privacy-Seeking Users Away from Reliable Information. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP '20 Adjunct)*. ACM, New York, NY, USA, 244–251.

[21] Ivan Srba, Robert Moro, Daniela Chuda, Maria Bielikova, Jakub Simko, Jakub Sevcech, Daniela Chuda, Pavol Navrat, and Maria Bielikova. 2019. Monant: Universal and Extensible Platform for Monitoring , Detection and Mitigation of Antisocial Behavior. In *Proceedings of Workshop on Reducing Online Misinformation Exposure (ROME 2019)*. 1–7.

[22] Siva Vaidhyanathan. 2018. *Antisocial media: How Facebook disconnects us and undermines democracy.* Oxford University Press.

[23] Xuezhi Wang, Cong Yu, Simon Baumgartner, and Flip Korn. 2018. Relevant Document Discovery for Fact-Checking Articles. In *Companion Proceedings of the The Web Conference 2018* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 525–533. https://doi.org/10.1145/3184558.3188723

[24] Q. Zhang, S. Liang, A. Lipani, Z. Ren, and E. Yilmaz. 2019. From Stances' Imbalance to Their Hierarchical Representation and Detection. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. ACM, New York, NY, USA, 2323–2332. https://doi.org/10.1145/3308558.3313724

[25] Zhe Zhao, Lichan Hong, Li Wei, Jilin Chen, Aniruddh Nath, Shawn Andrews, Aditee Kumthekar, Maheswaran Sathiamoorthy, Xinyang Yi, and Ed Chi. 2019. Recommending What Video to Watch next: A Multitask Ranking System. In *Proceedings of the 13th ACM Conference on Recommender Systems* (Copenhagen, Denmark) *(RecSys '19)*. ACM, New York, NY, USA, 43–51. https://doi.org/10.1145/3298689.3346997