Evaluating the Use of User Content Feed Swapping for Counteracting Filter Bubbles

Taylor Richmond, Lauri Tuovinen*

Biomimetics and Intelligent Systems Group, P.O. Box 4500, FI-90014 University of Oulu, Finland

Abstract

The term "filter bubble" refers to a phenomenon in which a social media recommendation system fails to offer diverse or novel content, and instead offers content that reinforces particular belief systems. Filter bubbles are considered harmful because of their potential polarizing effects in society and their role in the spread of false information online. In this paper, we propose a solution to counteract the effects of filter bubbles by providing users with the option to switch content feeds with their least similar user's feed. This is achieved by substituting the correlation coefficient used in collaborative filtering recommendation systems. A social media network simulation and accompanying questionnaire were used to test the viability of the solution. It was found to be viable in a simulated environment because it increased the users' self-reported bias perception, without adversely impacting user engagement metrics, after switching with their least similar user's feed. While a viable proof of concept in a simulated environment, the solution must be tested within a naturalistic setting with more participants in order to determine its real-world viability.

Keywords

Recommendation system, social media network, filter bubble, collaborative filtering, sentiment analysis, bias, perception, valence, content diversity

1. Introduction

Social media is an important tool for information dissemination and plays a significant role in shaping users' cognitive map of the world [1]. Recommendation systems are employed by social media networks (SMN) to provide relevant information to users. These systems optimize the scope of interest of a user through various metrics [2]. SMNs gather behavioral data to personalize recommendations, thereby enhancing the relevance and novelty of the content [2, 3, 4, 5, 6]. However, a phenomenon known as the filter bubble can occur when personalized recommendations become more relevant, less novel, and more filtered, leading to exposure to ideologically homogeneous content.

Social media platforms connect one third of the world's population and are a prime source of information for users [7, 8]. Thus, the content on a user's Facebook, Twitter, Instagram, and TikTok feed has immense political and social influence [9]. Social media has contributed to social change in countries like the United States, Canada, Iran, Pakistan, China, Egypt, Malaysia,

Conference on Technology Ethics - Tethics, October 18-19, 2023, Turku Finland

*Corresponding author.

taylor.richmond@oulu.fi (T. Richmond); lauri.tuovinen@oulu.fi (L. Tuovinen)

0000-0002-7916-0255 (L. Tuovinen)

and more [10]. It has played a crucial role in election campaigns, as seen in the 2008 American presidential elections and the 2009 Iranian presidential elections [10]. Social media has also enabled the spread of health misinformation [11, 12], pro-eating disorder narratives [13], anti-quarantine [14] and anti-vaccine movements [15, 16], hoaxes, propaganda, and disinformation in various contexts, including the 2018 Zimbabwe elections [17]. Filter bubbles have also been associated with the growth of many populist political movements, such as Brexit, Trump, and Bolsonaro [18].

The real-world impact of overly filtered content makes it important to counteract filter bubbles. Core human values are threatened if the polarization of political discourse escalates into violence or if exposure to health mis/disinformation leads people to engage in hazardous behavior, and while a direct causal relationship between the filter bubble phenomenon and outcomes such as these may be difficult to establish, from a utilitarian perspective filter bubbles should be counteracted if they are deemed likely to contribute to harm and if any negative consequences of the methods of counteraction are relatively mild. Besides avoiding harmful outcomes, there is also an opportunity to achieve beneficial outcomes by helping people expand their horizons in terms of content they enjoy, so counteracting filter bubbles is aligned with both beneficence and non-maleficence, arguably the two most important principles in technology ethics. Furthermore, exposing the existence and effects of filter bubbles to social media users can be viewed as promoting transparency, a key principle in the ethics of artificial intelligence.

This paper proposes a counteractive feature that targets filter bubbles and aims to expand users' cognitive maps by exposing them to the social media feeds of their least similar user. The proposed solution intends to achieve this goal without sacrificing user engagement. To test this concept, a simulated social media network utilizing a session-based collaborative recommendation system was developed. The system retrieves batches of recommendations by calculating user correlation scores and retrieving the most liked unseen posts of the most similar user. When the swap button is pressed, the same calculation occurs but with the least similar user instead.

Each post was colour-coded as either cool or warm. The simulation was tested by having ten participants choose an initial bias towards warm or cool tones. They were then immersed in a social media feed that was heavily biased towards their chosen colour tone, and then prompted to use the simulation as if they were browsing a real SMN. They could browse posts, read comments, and like posts. Before and after swapping feeds with their least similar user, they were given a series of questions to gauge their self-reported perception of how biased comments were in favour of or against different colour tones.

User engagement was not adversely affected by the swap. Both passive and active engagement metrics showed similar or increased engagement across user activity. The users' self-reported bias perception was impacted by the swap: they were more aware of bias, and when presented with a nominal scale to rate the level of bias perceived, they were more likely to choose nuanced answers following the swap than before. These results demonstrate the viability of the proposed solution as a proof of concept, but further research is needed in order to explore its viability in the real world.

The remainder of this paper is structured as follows: Section 2 discusses essential background and reviews related work. Section 3 describes the proposed solution and the protocol of the experiment by which it was tested. Section 4 presents the results of the experiment. Section 5

discusses the significance of the results as well as directions for future work. Section 6 concludes the paper.

2. Background and Related Work

2.1. Filter Bubbles

The term filter bubble was first coined by Eli Pariser and popularized in [19]. It refers to a unique information universe online wherein a user is predominantly exposed to ideologically similar content [1, 2, 3, 5, 20, 21, 22, 23, 24, 25, 26]. Each information universe is unique, as different users have different preferences and therefore the filter bubble will differ in appearance from user to user [6, 18, 27]. It should be noted that some academic studies reject the existence of filter bubbles; according to [28], individual preferences rather than algorithms determine users' exposure to "attitude challenging content" on Facebook. This suggests that users may choose to avoid content that conflicts with their beliefs, rather than algorithms being the primary cause of homogeneous content. Given the unknown influence of algorithms versus individual choice, this paper aims to minimize the potential impact of algorithms on individual choice.

Filter bubbles can be defined as a lack of information diversity. By exposing the user to more diverse content, and thus expanding that information diversity, the hypothesis is that their information universe will expand as well. In order to fully correct and counteract misinformation without backfiring, collaboration between computer scientists, psychologists, medical professionals, social scientists, and other professionals may be necessary [15]. Several attempts have been made across the years to both diagnose and counteract filter bubbles. In [23], a fairness criterion is proposed to determine whether inter-group links are represented in a link prediction algorithm output, as a means of determining whether group diversity is high or low, and thus diagnosing a potential filter bubble.

One proposed idea for counteracting filter bubbles is to maximize serendipity [2, 3], which can be defined as a mix of diversity, novelty, and relevance for recommended items, while not being heavily weighted towards any one or two of the three dimensions [3]. Another potential solution is increased awareness, which would involve designing new tools and techniques to encourage users to search for more diverse content; the solution proposed in this paper falls into this category. In [2], the use of visualizations was proposed to show users the categories of content they consume, and it was found that users had a better understanding of filtering through doing this.

2.2. Social Media Networks and Recommendation Systems

A key dimension to any research into social media and recommendation systems is the concept of engagement [7, 29]. In the real world, any change to a social media recommendation system is evaluated based on how well it achieves the goals of the SMN, and user engagement is a key goal [30]. User engagement keeps users on the site and using the service. For example, the measure of success of a YouTube algorithm is keeping the user on the site by having them watch an additional video after one video has finished [24]. Social media engagement comes in two forms: implicit and explicit feedback. Explicit feedback involves users reporting their interests,

such as by liking or commenting, whereas implicit feedback is obtained from observing user behaviour, represented by metrics such as dwell time, returns to a site, etc. [31].

Another way to define user engagement is to distinguish between passive and active engagement. Passive engagement is also called lurking, and is measured through delayed metrics such as the amount of time spent on posts and reading comments and the depth of post-viewing. Active engagement, in contrast, involves active participation, such as clicks, likes and comments [8, 32]. Traditionally, recommendation systems have been optimized for instant metrics, such as clicks, but more recently delayed metrics have also begun to be favoured [4]. Both active and passive engagement, as well as instant and delayed metrics are used to determine user engagement in this paper.

Recommendation systems take user feedback as input and provide a list of top N items the user is most likely to engage with using a variety of recommendation system techniques [33, 34, 35]. Widely used techniques include content-based filtering, collaborative filtering and hybrid filtering [35]. The system implemented for this paper uses collaborative filtering, which stems from leveraging collaborative behaviours of like-minded users to predict the behaviour of target users [33, 36, 37, 38].

As a user explores posts on an SMN, they form impressions and judgements regarding the content of those posts [39]. Comments on the posts can influence those judgements. One study found that when users read positive comments towards a company, their overall evaluation of that company was more positive. Conversely, one negative comment affected the company's reputation negatively [40]. In the experiment discussed in the following sections, comments are utilized as a means of revealing bias to users.

3. Implementation and Experiment Protocol

To explore the potential of feed swapping for counteracting filter bubbles, we devised a user study comprised of two parts: a SMN filter bubble simulation including a feed swap function, and an accompanying questionnaire. The SMN simulation was developed using a TKinter GUI, which displayed to the user, one by one, posts consisting of an image and a comment section. 200 images were generated for the posts using four online AI image generation tools: NightCafe, HotPot, Replicate, and Wombo. Wombo was used for the majority of image generations, as it was able to generate a high number of posts with different art styles, adding more variety to the content.

The generated posts were separated into two categories: 100 warm-toned posts and 100 cool-toned posts. A post was considered warm-toned if its RGBA red tones were double the value of the blue and green tones combined. A post was considered cool-toned if the RGBA blue tones were double the value of the red and green tones combined. These posts were then separated into two categories again, negative and positive, thus resulting in 50 positive warm-toned posts, 50 positive cool-toned posts, 50 negative warm-toned posts, and 50 negative cool-toned posts.

The polarity of the posts was determined through the comment section attached to each post. For this purpose, ChatGPT was used to generate hundreds of positive and negative comments. In order to perceive bias, language must be used that can be detected as biased

in one direction or another. ChatGPT was considered suitable for the task, since it is able to generate "grammatically flawless and seemingly-human replies to different types of questions and prompts" [41]. To eliminate ambiguity, negative posts had 100% negative comments, and positive posts had 100% positive comments.

There were 10 participants in the study, of whom 6 were aged 25+ and 4 were between the ages of 18 and 24. Before the start of the simulation, each user chose whether they preferred warm or cool tones. A simulated filter bubble was then created for the user by generating posts that were positive towards their preferred tone and negative towards their non-preferred tone. The posts were generated by calculating the most similar users to the current user and then choosing posts that they liked which the current user had not yet seen. These similar users were 40 simulated users with randomly generated likes matrices that were weighted towards a warm or cool bias.

The user could interact with the system by liking posts, checking comments and moving to next posts; additionally, they had the option to click on a feed swap button to see their least similar user's posts. These posts were generated by calculating which user had the least similar correlation to the current user by comparing their likes matrix, and then retrieving the most liked unseen posts of their least similar user. The calculation was done between the current participant's likes matrix and those of the 40 simulated users.

Post generation was done using a session-based, memory-based collaborative filtering technique. Session-based means that only interactions within a specific user session are taken into account when recommending, while memory-based means that the similarity scores of users are computed and stored in memory to produce new recommendations [42]. The statistical approach was chosen for this paper as there is no scalability issue due to recommendations being session-based. The statistical similarity measure used is Pearson Correlation, which is a common way of calculating similarity between users or items in recommendation systems [37]. The correlation is computed as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

The questionnaire was split into two parts, before and after the feed swap. The users were asked if they detected any bias in the posts they observed, and then requested to specify the nature of the bias. The questions were the same before and after the swap, except for the final question, which asked if the user preferred the posts before or after the swap. The full set of questionnaire questions is given in Appendix A.

4. Results of the Experiment

4.1. User Engagement

User engagement is not evenly distributed; some users engage much more than others [8]. Therefore, when we are evaluating user engagement, we will examine the average and standard deviation of user engagement for each user individually to determine if their engagement rose or fell, instead of taking a baseline average to compare their engagement to.

4.1.1. Active Engagement and Instant Metrics

Likes are the instant metric used to determine active engagement in this paper. Overall, users were active when engaging with posts. The user who liked the highest percentage of posts liked 42.31% of posts, while the user who liked the smallest number of posts liked 1.75%. Totalling all likes regardless of the colour tone of posts, the average number of likes was 23.4.

The average and standard deviation of likes was calculated by adding together the number of likes for negative cool-toned, positive cool-toned, negative warm-toned, and positive warm-toned posts. The results are presented in Table 1. From the table it can be seen that in 5 cases the number of likes increased after the swap, while in 5 cases the number decreased. That decrease was small in most cases, and resulted in less than half of the likes before the swap in only two cases.

From these statistics, it can be concluded that active engagement was adversely impacted by the swap in half of cases, but resulted in increased engagement in the other half. The decrease in engagement can be explained as an effect of the swap, but it may also be explained as a content saturation issue. Liking may be affected by how long an individual uses a site [8]; in other words, the longer users browse a SMN the less likely they are to actively engage with it.

Table 1
User likes statistics across differently toned posts

Participant	Before		After	
	Avg	Std	Avg	Std
A	3	2.74	6	4.3
B	8	7.4	4.5	5.45
C	1.25	1.89	2.25	2.87
D	4.75	5.19	1	1.41
E	3.25	3.4	2.25	2.87
F	1	2.37	2.25	2.22
G	0.25	0.5	0	0
H	3.75	3.3	6.75	7.27
I	4.75	2.27	4	4.24
J	0.5	0.58	0.75	0.96

4.1.2. Passive Engagement and Delayed Metrics

Two delayed metrics were used to evaluate passive engagement in the experiment: dwell time on posts and comment-related engagement such as how many times comments were checked and for how long. These user behaviour analytics were collected in the background while users browsed the simulated SMN. Overall, the highest number of AI-generated posts viewed was 237 and the lowest number was 29. The average number of posts viewed was 113.1, with a standard deviation of 66, indicating a high variance in the amount of content consumed. Of the 113 average posts viewed, users checked comments an average of 33.8 times. The individual participant results are presented in Table 2, where dt = average dwell time in seconds, cc = the

number of times comments were checked and ct = average amount of seconds comments were checked for.

From the table, it can be seen that 7 users spent more time on posts after the swap than before it. All users checked comments more after the swap, and 7 users spent less time on average checking comments after the swap. Since users were checking comments more frequently, it stands to reason that they would spend less time on more comments. The increase in dwell time, and increased interest in comments, suggests that after the swap the users displayed increased passive engagement towards the simulated SMN content.

It was hypothesized that the users may experience content fatigue after the swap and be less interested in content, resulting in decreased active engagement; however, this proved not to be the case. When delayed metrics are taken into account, user engagement increased after the swap in the majority of cases. It is possible that this increase was due to pure curiosity with regards to the user swap button and its consequences. For that reason, further research on the effects of the feedback, as well as changing back to the original user feed, would be required.

Table 2
User dwell time and comment statistics

Participant	Before			After		
	dt	cc	ct	dt	cc	ct
A	3.4	7	4.48	5	23	3.4
B	9.8	0	4.1	4.7	0	27.4
C	5.4	20	4.05	7	35	4.23
D	3	22	3.77	5.7	35	3.21
E	6.5	20	3.11	8.4	25	3.34
F	2.9	9	9.6	5.9	38	8.67
G	5.9	7	3.47	3.5	9	3.19
H	3.2	6	4.24	2.7	20	2.4
I	5.2	8	5.08	7.5	26	2.6
J	10.4	5	5.75	10.9	20	3.35

4.2. Expanding the User's Cognitive Map

The participants had a heightened awareness of bias after the swap. Before the content swap, only 6 out of the 10 participants reported detecting either a positive or negative bias towards either tone. After the swap, all participants detected a bias. There are several possible explanations for this result. One possibility is that prior to the swap, participants were not expecting to be asked about bias and therefore were not paying attention to it. Conversely, after the swap participants may have been actively seeking out bias in the content. It is also possible that the algorithm instantiating the change led to a more obvious bias. While no definitive conclusion as to why can be drawn from these results, it is clear that participants were able to detect bias in the majority of cases.

The participants had a more nuanced understanding of bias after the swap. Before the swap, in the majority of cases the participants detected an extreme bias, either extremely positive or extremely negative, whereas after the swap, there were more answers indicating detection of a

moderate or slight bias. Another result that points towards a shift in mentality towards bias is a shift in likes: after the swap the majority of participants switched preferences from their initial preference. In other words, if they chose a cool tone preference at the beginning, they liked more posts in favour of warm tones after the swap, and vice versa. The exact reason for this cannot be determined, but it is nonetheless a shift in preference and attention before and after the swap.

5. Discussion and future work

Overall, user engagement levels were unaffected by the feed swap, while user perception of bias was altered after the swap. Participants had a heightened awareness of bias post-swap, as evidenced by all ten participants detecting bias post-swap. They also had a more nuanced view of the bias, as their bias perceptions varied more after the swap compared to before. These findings show potential for content feed swapping to combat filter bubbles while not negatively affecting user engagement. Nonetheless, future investigations in a more naturalistic setting are necessary in order to determine the practicality of feed swapping.

This study included explicit disclosures of algorithmic modifications to the users, which could have influenced their receptivity to it. In a naturalistic settings, users may resist algorithmic changes [43]. Making the content feed swap voluntary is one way of potentially avoiding such resistance.

This solution is meant to serve as a foundation for future studies aimed at counteracting filter bubbles. In order to establish its practicality, it must be tested on a larger user base in an uncontrolled, real-world environment. The small sample size of this study means that its results may be anomalous. Equal opportunity and statistical parity are of great importance when determining bias in a recommendation system [2]; these cannot be ensured with such a small participant pool. In addition, the solution must be extended to include a latent model-based collaborative recommendation system, wherein correlation neighbourhoods are inverted to calculate the least correlated items. As a result of these limitations, the results of this study cannot be generalized and can only act as a foundation for a study with a larger participant pool and which implements a model-based collaborative recommendation system.

6. Conclusion

This paper presented a novel approach to promoting content diversity and increasing awareness of content bias in recommendation systems by counteracting the effects of filter bubbles. The proposed solution involved providing users with the option to swap content feeds with their least similar user's feed. This was done by substituting the correlation coefficient used in a collaborative filtering recommendation system. The viability of this solution was tested utilizing a social media network simulation and accompanying questionnaire.

The solution was determined to be viable as a proof of concept, since it led to an increase in users' self-reported bias perception without adversely impacting user engagement metrics. When users reported their perception of bias in the content, their reports became more nuanced following the swap. Meanwhile, engagement metrics such as likes, dwell time, and time spent

checking comments did not decrease after the swap. These results show potential, but to assess the real-world viability of the proposed solution, it must be tested with a larger user base in a naturalistic setting.

References

[1] Z. Sawicka, How Facebook polarizes public debate in Poland-Polish filter bubble, Social Communication 5 (2019) 45–52.

[2] S. Nagulendra, J. Vassileva, Understanding and controlling the filter bubble through interactive visualization: A user study, HT 2014 - Proceedings of the 25th ACM Conference on Hypertext and Social Media (2014) 107–115. doi:10.1145/2631775.2631811.

[3] V. Maccatrozzo, Burst the filter bubble: Using semantic web to enable serendipity, in: The Semantic Web – ISWC 2012, 2012, p. 391–398. doi:10.1007/978-3-642-35173-0_28.

[4] L. Zou, J. Song, L. Xia, W. Liu, Z. Ding, D. Yin, Reinforcement learning to optimize long-term user engagement in recommender systems, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2019) 2810–2818. doi:10.1145/3292500.3330668.

[5] T. T. Nguyen, P.-M. Hui, F. M. Harper, L. Terveen, J. A. Konstan, Exploring the filter bubble: The effect of using recommender systems on content diversity, in: Proceedings of the 23rd international conference on World wide web, 2014, pp. 677–686.

[6] P. Symeonidis, L. Coba, M. Zanker, Counteracting the filter bubble in recommender systems: Novelty-aware matrix factorization, Intelligenza Artificiale 13 (2019) 37–47. doi:10.3233/IA-190017.

[7] R. Dolan, J. Conduit, J. Fahy, S. Goodman, Social media engagement behaviour: A uses and gratifications perspective, Journal of strategic marketing 24 (2016) 261–277.

[8] M. L. Khan, Social media engagement: What motivates user participation and consumption on youtube?, Computers in Human Behavior 66 (2017) 236–247. doi:10.1016/j.chb.2016.09.024.

[9] L. Bode, Political news in the news feed: Learning politics from social media, Mass Communication and Society 19 (2016) 24–48. doi:10.1080/15205436.2015.1045149.

[10] A. M. Attia, N. Aziz, B. Friedman, M. F. Elhusseiny, Commentary: The impact of social networking tools on political change in Egypt's "revolution 2.0", Electronic Commerce Research and Applications 10 (2011) 369–374. doi:10.1016/j.elerap.2011.05.003.

[11] S. Johnson, M. Parsons, T. Dorff, M. S. Moran, J. H. Ward, S. A. Cohen, W. Akerley, J. Bauman, J. Hubbard, D. E. Spratt, C. L. Bylund, B. Swire-Thompson, T. Onega, L. D. Scherer, J. Tward, A. Fagerlin, Cancer misinformation and harmful information on Facebook and other social media: A brief report, Journal of the National Cancer Institute 114 (2022) 1036–1039. doi:10.1093/jnci/djab141.

[12] A. L. Svalastog, J. Allgaier, S. Gajovic, Navigating knowledge landscapes: On health, science, communication, media, and society, Croatian Medical Journal 56 (2015) 321–333. doi:10.3325/cmj.2015.56.321.

[13] V. Suarez-Lledo, J. Alvarez-Galvez, Prevalence of health misinformation on social media: Systematic review, Journal of medical Internet research 23 (2021) e17187.

[14] A. Karami, M. Anderson, Social media and COVID-19: Characterizing anti-quarantine comments on Twitter, Proceedings of the Association for Information Science and Technology 57 (2020) e349.
[15] Y. Wang, M. McKee, A. Torbica, D. Stuckler, Systematic literature review on the spread of health-related misinformation on social media, Social science & medicine 240 (2019) 112552.
[16] M. C. Jenkins, M. A. Moreno, Vaccination discussion among parents on social media: A content analysis of comments on parenting blogs, Journal of Health Communication 25 (2020) 232–242. doi:10.1080/10810730.2020.1737761.
[17] M. N. Ndlela, W. Mano, Social media and elections in Africa, volume 1: Theoretical perspectives and election campaigns, Springer Nature, 2020.
[18] A. Bruns, Filter bubble, Internet Policy Review 8 (2019). doi:10.14763/2019.4.1426.
[19] E. Pariser, The Filter Bubble: What the Internet Is Hiding from You, Penguin Books, 2011.
[20] P. Seargeant, C. Tagg, Social media and the future of open debate: A user-oriented approach to Facebook's filter bubble conundrum, Discourse, Context & Media 27 (2019) 41–48.
[21] A. Bechmann, K. L. Nielbo, Are we exposed to the same "news" in the news feed? An empirical analysis of filter bubbles as information similarity for Danish Facebook users, Digital journalism 6 (2018) 990–1002.
[22] T. Graham, R. Ackland, Do socialbots dream of popping the filter bubble? The role of socialbots in promoting deliberative democracy in social media, Socialbots and their friends: Digital media and the automation of sociality (2017) 187–206.
[23] F. Masrour, T. Wilson, H. Yan, P.-N. Tan, A.-H. Esfahanian, Bursting the filter bubble: Fairness-aware network link prediction, Proceedings of the AAAI Conference on Artificial Intelligence 34 (2020) 841–848. doi:10.1609/aaai.v34i01.5429.
[24] L. V. Bryant, The YouTube algorithm and the alt-right filter bubble, Open Information Science 4 (2020) 85–90. doi:10.1515/opis-2020-0007.
[25] G. Iannelli, G. De Marzo, C. Castellano, Filter bubble effect in the multistate voter model, Chaos: An Interdisciplinary Journal of Nonlinear Science 32 (2022) 043103.
[26] R. Barker, Trapped in the filter bubble? Exploring the influence of Google search on the creative process, Journal of Interactive Advertising 18 (2018) 85–95. doi:10.1080/15252019.2018.1487810.
[27] D. O'Callaghan, D. Greene, M. Conway, J. Carthy, P. Cunningham, The extreme right filter bubble, arXiv preprint:1308.6149 (2013).
[28] A. Pippin, Social Media and the Filter Bubble: Curated Flows Theory, Facebook, and News Diversity, The University of Alabama, 2022.
[29] L. McCay-Peet, A. Quan-Haase, A model of social media engagement: User profiles, gratifications, and experiences, Springer, 2016.
[30] G. Shani, A. Gunawardana, Evaluating recommendation systems, Recommender systems handbook (2011) 257–297.
[31] Y. Koren, Factorization meets the neighborhood: A multifaceted collaborative filtering model, in: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008, pp. 426–434.
[32] H. Shahbaznezhad, R. Dolan, M. Rashidirad, The role of social media content format and platform in users' engagement behavior, Journal of Interactive Marketing 53 (2021) 47–65.

doi:10.1016/j.intmar.2020.05.001.
[33] F. O. Isinkaye, Y. O. Folajimi, B. A. Ojokoh, Recommendation systems: Principles, methods and evaluation, Egyptian Informatics Journal 16 (2015) 261–273. doi:10.1016/j.eij.2015.06.005.
[34] M. Etter, O. B. Albu, Activists in the dark: Social media algorithms and collective action in two social movement organizations, Organization 28 (2021) 68–91.
[35] U. Javed, K. Shaukat, I. A. Hameed, F. Iqbal, T. M. Alam, S. Luo, A review of content-based and context-based recommendation systems, International Journal of Emerging Technologies in Learning 16 (2021) 274–306. doi:10.3991/ijet.v16i03.18851.
[36] L. Wu, X. He, X. Wang, K. Zhang, M. Wang, A survey on accuracy-oriented neural recommendation: From collaborative filtering to information-rich recommendation, IEEE Transactions on Knowledge and Data Engineering 35 (2022) 4425-4445.
[37] D. Xu, C. Ruan, E. Korpeoglu, S. Kumar, K. Achan, Rethinking neural vs. matrix-factorization collaborative filtering: The theoretical perspectives, in: International Conference on Machine Learning, PMLR, 2021, pp. 11514–11524.
[38] S. Dhelim, N. Aung, M. A. Bouras, H. Ning, E. Cambria, A survey on personality-aware recommendation systems, Artificial Intelligence Review (2022) 2409–2454.
[39] C. Edwards, A. Edwards, P. R. Spence, A. K. Shelton, Is that a bot running the social media feed? Testing the differences in perceptions of communication quality for a human agent and a bot agent on Twitter, Computers in Human Behavior 33 (2014) 372–376.
[40] M. J. Lee, J. W. Chun, Reading others' comments and public opinion poll results on social media: Social judgment and spiral of empowerment, Computers in Human Behavior 65 (2016) 479–487.
[41] S. Mitrović, D. Andreoletti, O. Ayoub, ChatGPT or human? detect and explain. Explaining decisions of machine learning model for detecting short ChatGPT-generated text, arXiv preprint:2301.13852 (2023).
[42] M. Marcuzzo, A. Zangari, A. Albarelli, A. Gasparetto, Recommendation systems: An insight into current development and future research challenges, IEEE Access 10 (2022) 86578–86623. doi:10.1109/ACCESS.2022.3194536.
[43] M. A. D. Vito, D. Gergle, J. Birnholtz, "Algorithms ruin everything": #RIPTwitter, folk theories, and resistance to algorithmic change in social media, Conference on Human Factors in Computing Systems - Proceedings 2017-May (2017) 3163–3174. doi:10.1145/3025453.3025659.

A. Questionnaire Questions

1. Age
Options: 18-24 | 25+ | Prefer not to say

2. Thinking about the posts and comments you observed, did you notice any bias towards cool colours or warm colours?
Options: Yes | No

If yes, was the bias towards cool or warm colours? How positive or negative was the observed bias? Answer the following multiple choice grid based on your observations:
Cool
Options: Very Positive | Moderately Positive | Slightly Positive | Neutral | Slightly Negative | Moderately Negative | Very Negative
Warm
Options: Very Positive | Moderately Positive | Slightly Positive | Neutral | Slightly Negative | Moderately Negative | Very Negative

3. Did you choose to press the ”swap” button?
Options: Yes | No

4. Thinking about the posts and comments you observed, did you notice any bias towards cool colours or warm colours?
Options: Yes | No

If yes, was the bias towards cool or warm colours? How positive or negative was the observed bias? Answer the following multiple choice grid based on your observations:
Cool
Options: Very Positive | Moderately Positive | Slightly Positive | Neutral | Slightly Negative | Moderately Negative | Very Negative
Warm
Options: Very Positive | Moderately Positive | Slightly Positive | Neutral | Slightly Negative | Moderately Negative | Very Negative

5. Did you prefer the content before or after the swap?
Options: Before | After | No preference