Taking Back Control of Social Media Feeds with *Take Back Control*

Joni Salminen
Qatar Computing Research Institute
Hamad Bin Khalifa University
Turku School of Economics
Turku University
Doha, Qatar; Turku, Finland
jsalminen@hbku.edu.qa

Juan Corporan
Banco Santa Cruz RD
Dominican Republic
juan.nunez.corp@gmail.com

Soon-gyo Jung
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
sjung@hbku.edu.qa

Bernard J. Jansen
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
jjansen@acm.org

***Abstract*— Controlling the quality of social media feeds poses an issue for many users. Platforms such as Twitter give users some options to influence their feeds. Still, the selection of content predominantly relies on implicit rather than explicit user actions, as manual options for "cleaning the feed" are often cumbersome and difficult to use for most users. Here, we present *Take Back Control*, a web browser extension that gives users control to hide undesirable content from their social media feeds. The extension combines JavaScript (for hiding the content) and machine learning (for deciding what content to hide). Our current demonstration includes three filter types: Toxic, Political, and Negative content, with a possibility to add more filters, all of this with the overarching aim of helping end users control the information visible in their social media feeds.**

Keywords*— *Social Media, User Control, Feeds, Browser Extension

I. Introduction

Social media has transformed the way many people receive their information about the world [47]. This information includes updates from friends and family, professional content that helps improve knowledge and skills for modern knowledge-intensive trades, and news content that keeps users updated on what is happening worldwide. While social media has enhanced information dissemination [42], several challenges are associated with social media. Examples of these challenges are online toxicity, hate speech, fake news, and general negativity [8,24,33,35,43]. Some users might, for example, want to turn off news about COVID-19, as this content may cause anxiety to them or block political content during elections.

While users might want to avoid specific types of content, the social media platforms are not necessarily giving them effective tools for doing so. For example, the platforms typically have no option to "turn off political posts", even though overly politicized news feeds may cause users stress, anxiety, and even anger. Furthermore, certain types of content may trigger societal adverse effects, such as group polarization, loss of focus/productivity, addiction, and mental health issues [11,12,48]. There is a need to give users better tools to control the content being displayed on their feeds. Social media platforms such as Facebook and Twitter are aware of this issue and have added functionalities towards resolving it. For example, Facebook enables one to unfollow pages and people and hide content from certain sources. Twitter has a keyword-based functionality to prevent the showing of content with specific terms and hashtags[1].

However, for the most part, the platforms still rely on *implicit* control (i.e., the newsfeed algorithm) instead of *explicit* control (i.e., hard rules from users of what type of content they want to see). Our premise is that explicit controls are needed, and the computer science community, as a frontrunner in users' interaction with social media, should investigate such techniques.

To this end, this study proposes an approach towards moving from implicit control of feeds to explicit control. By explicit, we imply giving the user more precise and direct control over the content in their feeds. In the implicit model, the social media platform's algorithm determines the type of content one sees based on clicks, likes, comments, the pages, and people followed, and so on. These implicit actions, however, may not always correspond to the user's needs. Consider an example of seeing something that the user dislikes — it may be that the user does not hide or ignore that content but argues against it by writing a comment. This would indicate to the algorithm that the user *is* interested in this content, whereas, in reality, they would prefer not to see it.

This example illustrates the problem of users lacking control over the content they see and the fact that the lack of this control leads to many unwanted consequences. Given the rise in politicization, polarization, dissatisfaction, and negative news on social media, some people would prefer to stay shielded from such content, while still remaining users of different social media platforms so they can engage with other types of content. Developing tools for users to take back control over their social media feeds is relevant for reducing negative effects associated with the use of social media, such as anxiety and stress [11].

With this in mind, this research reports the development and implementation of **Take Back Control** (TBC), an academic-driven tool that uses machine learning (ML) models to hide specific types of content from a user's feeds based on the user's explicit needs.

II. Related Work

Research has reported numerous negative psychological effects associated with the use of social media, such as stress, anxiety, feelings of inferiority, depression, and anger [15,18,23,27,32,45,50]. Typically, these are direct or indirect consequences from exposure to specific types of online

[1] https://help.twitter.com/en/using-twitter/advanced-twitter-mute-options

978-1-6654-3603-8/21/$31.00 ©2021 IEEE

content. While users' development of cognitive strategies to cope with harmful content offer one way to mitigate these effects [40], the exposure to undesired content remains a challenge for the psychological wellbeing of online users.

There is little extant work on controlling the content in social media feeds using computational tools and techniques. Perhaps the closest stream of research originates from the explicit and implicit signals in recommendation systems. Social media feeds can be viewed as recommender systems [4], relying on implicit and explicit cues. Implicit cues are based on behaviors such as scrolling, viewing, and hovering [30]. Explicit cues are elaborate actions, such as clicking, liking, commenting, and sharing [22]. Explicit control is also gained through options, which are features built into the social media platform to give users control over their feeds [21]. Users may not always be aware of these options, find them, or effectively use them [21,44].

Overall, it is doubtful if the platforms are doing everything in their power to help users gain the ultimate control over their feeds. One reason to doubt this is because the platforms have a vested interest in not maximizing the happiness or contentment of their users but the time spent on the platform, as the platforms monetize this time by showing online ads [51]. As such, the research community should not rely solely on the platforms to build tools for changing the content on social media feeds.

Another issue with implicit control is that it gives little control over the sentiment or category of the content. The user can unfollow pages or people, but they cannot unfollow broader themes like 'politics'. This is a problem because users' posts are a mixture of themes [5,31,41] — for example, a data scientist professional may post about data science related topics 80% of the time and about politics for the rest of the 20%. However, the user following this person might only be interested in the data science content that the individual posts. Yet, the only way to avoid being exposed to the political content posted by the individual is to unfollow him, which would make the interested person lose out on the data science related content.

Our research addresses this need for academic-driven tools that give users greater control over the type of content they see in social media. An analogy can be found in ad blockers [1,2,7], i.e., browser extensions that hide ads (=undesirable content) on web pages.

III. Methodology

Here, we explain the logic and functionality of TBC, which is implemented in this demonstration as a web browser extension for Google Chrome that controls Twitter feeds. Chrome was selected because of its wide user base and open programming specifications that enable users to run local and centralized ("marketplace") extensions to enhance their browsing experience. Twitter was selected for similar reasons; note that the ideas proposed here could be applied to other browsers (e.g., Firefox, Edge) and platforms (e.g., Facebook, YouTube) as well, since they provide similar applicability from a development standpoint.

A. Design Principles

We defined four design principles for the extension:

- **Transparency**: The extension informs the user of the choices it makes.
- **Accuracy**: The extension gives accurate results.
- **User-friendliness**: The extension gives fast results without compromising users' overall browsing experience on social media.
- **Openness**: Others can build on the work presented.

For transparency, we built a logging system that enables users to view the tweets sent to the API and the returned labels for each. This enables evaluation of the model correctness and tells the user what tweets the models considered negative/toxic/political.

For speed, we consider it as an evaluation criterion for the choice of the ML algorithm by choosing the algorithm with the best balance across speed and accuracy. We use asynchronous web development techniques (Ajax) to mitigate interference with the display of social media posts for the client-side.

For accuracy, we choose ML models that provide satisfactory performance when tested on unseen test data. Particularly, a high false positive rate means you would see many feeds never making it to the timeline. A high false negative rate means the extension would be ineffective.

For openness, we make the extension's source code publicly available; report its development to enable replication, and develop the architecture to support the expansion of the extension by adding filters (POST requests to the API asking for an updated list of filters).

B. Functionality

The user can enable/disable any of the available filter types in the tool's user interface (UI). Based on this choice, the tool will hide the corresponding tweets from the user's feed (see Fig. 1). We demonstrate the application using three filter types (Politics, Toxicity, Negativity) that are ML models specifically developed for this work and deployed via an application programming interface (API). Other filter types can be added to TBC, as its server-side and client-side are separate modules (see Fig. 2). The extension retrieves an updated list of filters from the API (server) at each startup and updates the UI (client browser) accordingly.

The extension reads the tweets' text content while the user browses the feed, sends the text to the API, and waits for the prediction result (see Fig. 1). As the user scrolls their feed, new content emerges. This content is instantaneously sent to the API. The tweets corresponding to the filters the user has enabled are hidden by using JavaScript. The current implementation is fast enough so that the latency of getting the predictions should not hinder the end-user experience of scrolling the feed (see evaluation in Section 4).

C. Machine Learning Filters

In this section, we report the ML models used for filtering. We will use datasets that contain short texts from each filter type (political, toxic, and negative content) to train binary classifiers. Each filter has its own classifier. To evaluate the performance of the models, we use three metrics: (a) F1 score that is the harmonic mean of precision and recall [34], (b) prediction speed (S) of the model in seconds, and (c) Aggregate Score (A), a metric that indicates the average performance difference between algorithms *i* and *j*:

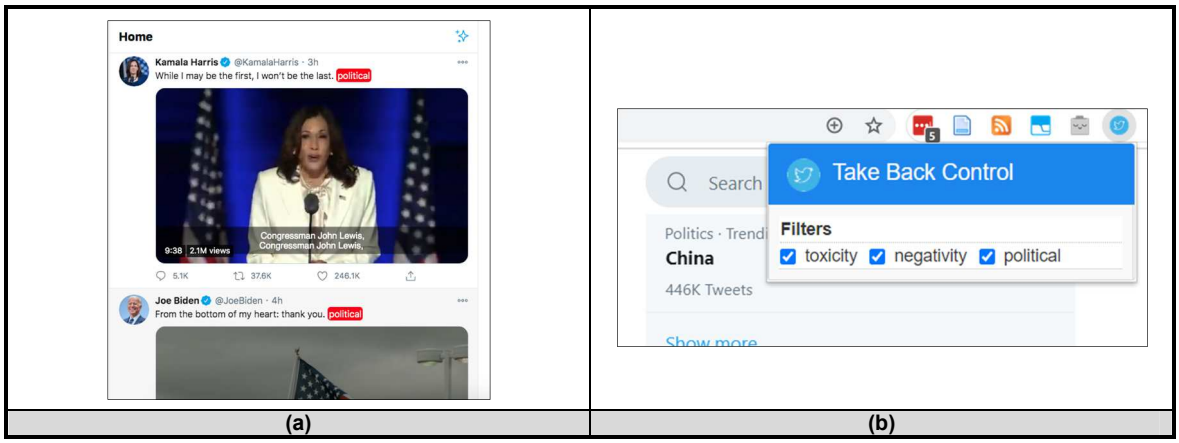

Fig. 1. TBC's option dialogue (b). If the filter 'political' is enabled and a tweet is classified political, the extension will automatically hide it (labels in (a) are shown for illustration).

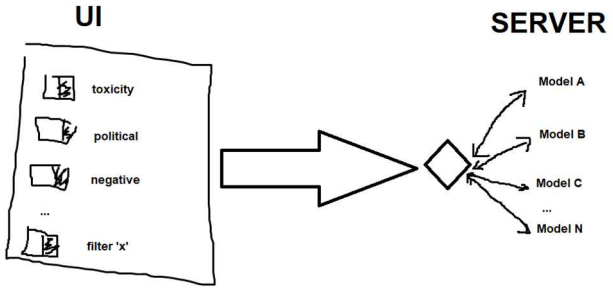

Fig. 2. The design of the 'Take Back Control' extension. The extension sends tweets to the API to analyze if they contain unwanted content. The extension will only ask for an output from a specific model if the user has enabled the corresponding filter in the extension UI. The positive cases are hidden from the user's social media feed.

$$A_i = \frac{\frac{F1_i - F1_j}{F1_i} + \frac{1/S_i - 1/S_j}{1/S_i}}{2}$$

In other words, we presume that the improvement/worsening of speed is equally important as the improvement/worsening of the F1 score because a slow prediction would hamper the user experience and require the user to take pauses while scrolling their feeds. We use the A metric for the final choice of models since the models need to be fast and reliable to offer a pleasant user experience.

We compare a Support Vector Machine (SVM) and the DistilBERT [38] transformer for algorithms. SVM is considered a responsive classifier with decent all-around performance [29]. DistilBERT is a state-of-the-art NLP model based on BERT, a recent breakthrough [14] in NLP. We use the 'distilled' version, which contains less complexity (fewer parameters) than the original BERT model but obtains very similar performance [38].

We collected data using publicly available datasets from Kaggle, an online community of data scientists and ML practitioners that contains datasets from numerous sources for training our ML models.

- **Dataset 1:** News category dataset from Kaggle[2]. There are 41 news categories. We use the 32,739 samples in the politics and merge the other 40 classes to obtain 168,114 non-political samples.
- **Dataset 2:** Toxicity dataset from Kaggle[3]. Contains 16,225 samples of toxic comments, and 143,346 samples of neutral (non-toxic) comments.
- **Dataset 3:** Sentiment dataset from Kaggle[4]. Contains 800,000 samples of the positive class and 800,000 samples of the negative class.

TABLE I. EXPERIMENTAL RESULTS OF THE TESTED MODELS.

Politics					
SVM			DistilBERT		
F1	S	A	F1	S	A
0.85	0.06	**0.46**	0.91	6.02	-47
Toxicity					
SVM			DistilBERT		
F1	S	A	F1	S	A
0.86	0.12	**0.45**	0.94	7.18	-29
Negativity					
SVM			DistilBERT		
F1	S	A	F1	S	A
0.67	0.03	**0.39**	0.81	9.21	-153

Note: Speed (S) indicates the average prediction time in seconds, when predicting 1,000 samples across seven runs and ten loops.

For preprocessing the features, we use HuggingFace's [49] function that preprocesses the text into the format expected by the pretrained model. Preprocessing for SVM involved

[2] https://www.kaggle.com/rmisra/news-category-dataset
[3] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data
[4] https://www.kaggle.com/kazanova/sentiment140

removing URLs, extra spaces, common English stop words, and non-letter symbols, as well as converting the tokens to lower case. The tokens (words) in the training set were transformed into a matrix of token counts for the Linear Support Vector Classifier that the SVM implementation uses. The train/test split was 80/20 for the comparative experiments reported in Table 1, and we used a stratified k-fold of k=7 for the optimized final models. The hyperparameters of the final models were optimized using grid search [28] before deploying them via the API.

From the results in Table 1, we can make two observations. In terms of the F1 score, DistilBERT is always better, and in terms of speed, SVM is always better. Since the relative differences are much greater for speed than for the F1 score, the aggregate scores favor SVM for each classification task. Therefore, we implement the SVM classifiers for the extension.

IV. Evaluation

We initially tested the extension with Google Chrome Version 86.0.4240.183 (Official Build) (64-bit). Installation is simple: one needs to go to chrome://extensions from the browser, ensure that the Developer mode is enabled, click 'Load unpacked', and select the folder in which the extension's zip file is unpacked. After this, the TBC extension shows up in the browser. Clicking its icon on Chrome's top-right toolbar will reveal the tool UI, and after this, the Twitter feed should be refreshed (F5 key in the browser). The filtering should now be operational.

Three users installed the extension and enabled all the filters while browsing Twitter for a typical session of 5-10 minutes, in which they were asked to browse their Twitter feeds naturally. After browsing, the users downloaded their log files that stored tweets sent to the server, and the responses (yes/no) received for each class. We subsequently analyzed these logs. Table 2 summarizes the results of the user sessions.

TABLE II. Summary statistics of the user testing.

	User 1	**User 2**	**User 3**
	Male, 35, Finland	***Male, 36, Korea***	***Male, 60, USA***
Total tweets evaluated	n=175	n=207	n=131
Total tweets hidden (% of total)	39 (22.3%)	138 (66.7%)	52 (39.7%)
Politics (% of hidden)	6 (15.4%)	46 (33.3%)	13 (25.0%)
Toxicity (%)	3 (7.7%)	9 (6.5%)	2 (3.8%)
Negative (%)	30 (76.9%)	83 (60.1%)	37 (71.2%)
Note: Percentage (%) indicates the share of positive cases from the total evaluated cases.			

In a test session, User 1 (Male, 35, Finland) browsed his Twitter feed naturalistically for a duration of approximately five minutes with the TBC extension enabled. In this time, the extension evaluated 175 tweets and hid 39 (22.3%) of them (see Table 2, User 1 column). The majority of the hidden tweets (n=30, 76.9%) were hidden based on negativity. Six tweets (15.4%) were hidden due to political content, and three tweets (7.7%) because TBC believed them to contain toxic content.

Results in Table 2 indicate that there are major differences by user. For example, based on a Chi-squared analysis, TBC hides significantly more tweets from User 2 than User 1, X^2 (1, N = 382) = 75.1, $p < 0.001$. This can be seen as a natural consequence of the very different content compositions among the feeds. One pattern, however, is that negative tweets are more common than other tweet types, and the rank of filtered content is the same for all users: *negative*, then *political*, and then *toxic*. This seems intuitively logical, as negative content is more prevalent than political (as politics is a specific topic domain), and toxicity should be a relatively rare instance in one's social media feed, since one self-selects which accounts to follow.

For User 2, when all filters are enabled, the tool hides most (66.7%) tweets. This is due to the issue of false positives, which we will discuss shortly. However, when disabling the negativity filter, the rate of filtering drops to 23.7% (n=49 hidden tweets). This is still a high number but considerably lower than when all filters are enabled. User 3's ranking of percentages was the same as for the other users. The macro-average hide rate for all three users was 42.9%.

A relatively small number of instances were labeled for more than one condition (see Table 3). For example, the tweet "Many conservatives fear climate change mitigation, universal health care, and the protection of democratic rights much more than literal fascism". (from User 1's feed) was labeled as both political and negative. The tweet "so b/c of discussion about how quickly Elliot Page's Wikipedia page changed (not surprising, of course!) I checked out their gender identity policy, and Like, it's full of FAQs that are basically 'stop asking stupid questions like this one:"' (User 2's feed) was labeled as toxic and negative. The tweet "On its 50th anniversary, the US Environmental Protection Agency has been defanged after four years of Trump deregulation, but Biden has pledged to reinvigorate the government's guardian of Mother Nature." (User 2's feed) was labeled as toxic and political.

TABLE III. The number of tweets labeled with more than one filtering condition.

	User 1	**User 2**	**User 3**
Political AND Negative	3 (1.7%)	19 (9.2%)	3 (2.3%)
Toxic AND Negative	2 (1.1%)	3 (1.4%)	1 (0.8%)
Toxic AND Political	0 (0%)	6 (2.9%)	1 (0.8%)
Toxic AND Political AND Negative	0 (0%)	0 (0%	0 (0%)
Note: Percentage (%) indicates the share of total evaluated tweets for the user.			

Finally, based on the qualitative user observation, there was no lag in the scrolling when enabling the extension. This is to be expected as the extension does not control the rendering of the page but only manipulates the content after it has been rendered.

V. Limitations and Future Work

A closer look into the classified content (see Fig. 3) shows that the models do commit mistakes – instances that are not political/toxic/negative are labeled as such (false positives) and instances that are of those types are not considered as such (false negative).

Fig. 3. Examples of false positives.

There are many possible root causes to this. First, the training sets can be incomplete, i.e., not covering all the instances in the real world. Second, the distributions of words and expressions in the training set may not adequately compare to the use of language in the wild [39], as the training set is static and language in the wild is continually evolving. Third, no matter how good an ML model is, there exists room for errors. Tasks such as toxicity detection and sentiment analysis have shown to be difficult [9,10,19,35,37,46], not only for machines but also for humans due to subjectivity and interpretation.

Particularly, the positive/negative test set may be *more* balanced than in the wild (i.e., contain more negativity than what is natural), resulting in the model 'seeing negativity everywhere'. While balanced datasets are often good for research purposes, it would be worthwhile to build datasets that correspond with the real data distributions on users' feeds for a production system such as this. This is an exciting and particularly challenging area for further investigation.

This work presented a TBC prototype with basic functionality to filter content based on user needs. Several interesting avenues remain for future research and development:

- **Additional filters.** Others could build upon this work and introduce their own filters. There could be a "marketplace" of filters where developers could describe the type of filter, and users could rate the filters and share their comments on their behavior.
- **Multilingual compatibility.** Currently, the three models only support English. Community research effort is needed to cover more languages.
- **Mobile app compatibility.** The approach of using an extension only works with devices whose browsers support extensions (desktops, laptops). This excludes mobile devices – therefore, techniques for increasing user control in mobile devices warrant further research.
- **Improving the quality of the filters.** This requires training new models using more diverse training sets or, applying online learning, in which the users could fix the incorrect labels as they appear. These would then be sent back to the API to retrain the model.
- **User studies.** Studies on what filters users would select, why they select them, and how applying the filters affects their overall experience on social media, related to the on-going trends of "HCI for good", explainability, and transparency of algorithmic decision making [25,26], and could enhance our understanding of social media users' interactions with social media platforms.

VI. Discussion

Over-politicization, toxicity, and negativity degrade user experience on social media. Rather than censoring content on behalf of users, we promote an approach of giving users more control over the content they see. This control can come in the form of *filters*, as demonstrated in our implementation of the TBC browser extension. Using this filter-based approach, users could opt into political content, for example, then opt-out for a time to 'take a break', and then opt back in, all with ease. This manuscript proposes a starting point. Although ours was a prototype of what could be possible, we see that continuing this line of work, with effort from the community, would establish a positive impact, empowering users and giving them more control over their social media experience [3]. All resources, including the applied ML models and the browser extension, along with its source code, are shared in the Supplementary Material[5].

An interesting side discussion concerning the theme of "taking back control" is that one can legitimately ask: "Did we ever have control?" In the era of mass media, the information was filtered by journalists and sometimes by political parties. At the dawn of social media, users were led to believe that algorithms do all the work to show us the information we need to see. This, of course, turned out not to be the case, as studies in algorithmic bias are currently demonstrating at an increasing pace [6,16,20,36]. Therefore, one can claim that there never was a time when the commoner was in control of the information. Nevertheless, the essential question is: *could* we be in control? Technically, we have the tools for that. The information age enables the free flow of information — if properly structured. It is fascinating that despite decentralization and user-generated content, the power of deciding on the compositions of users' feeds is centralized to so few parties. As it stands, one can claim that the social media platforms command too much power over information diffusion [12,13,17,21,44].

VII. Conclusion

Users spend millions of hours on social media feeds every day. Therefore, improving the quality of the content on these feeds is an important mission for researchers and practitioners alike. Our academic-driven tool demonstrates an approach of giving users more control over their feeds – we hope that this example would inspire community effort around this crucial topic of letting users 'take back control'.

5 https://www.dropbox.com/sh/pvsto416dat5iex/AADM5Y0ttsJVQonSy5FgA_xva?dl=0

REFERENCES

[1] Maximilian Altmeyer, Kathrin Dernbecher, Vladislav Hnatovskiy, Marc Schubhan, Pascal Lessel, and Antonio Krüger. 2019. Gamified Ads: Bridging the Gap Between User Enjoyment and the Effectiveness of Online Ads. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

[2] Maximilian Altmeyer, Pascal Lessel, Kathrin Dernbecher, Vladislav Hnatovskiy, Marc Schubhan, and Antonio Krüger. 2019. Eating Ads With a Monster: Introducing a Gamified Ad Blocker. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–6.

[3] Oscar Alvarado and Annika Waern. 2018. Towards algorithmic experience: Initial efforts for social media contexts. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12.

[4] Anitha Anandhan, Liyana Shuib, Maizatul Akmar Ismail, and Ghulam Mujtaba. 2018. Social media recommender systems: review and open research issues. *IEEE Access* 6, (2018), 15608–15628.

[5] Annika Bergström and Maria Jervelycke Belfrage. 2018. News in social media: Incidental consumption and the role of opinion leaders. *Digital Journalism* 6, 5 (2018), 583–598.

[6] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, 4349–4357. Retrieved March 21, 2017 from http://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings

[7] Henriette Cramer. 2015. Effects of ad quality & content-relevance on perceived content quality. In *proceedings of the 33rd annual ACM conference on human factors in computing systems*, 2231–2234.

[8] Laura Dabbish, Gloria Mark, and Víctor M. González. 2011. Why do I keep interrupting myself? Environment, habit and self-interruption. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3127–3130.

[9] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, Association for Computational Linguistics, Florence, Italy, 25–35. DOI:https://doi.org/10.18653/v1/W19-3504

[10] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of Eleventh International AAAI Conference on Web and Social Media*, Montreal, Canada, 512–515.

[11] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2098–2110.

[12] Michela Del Vicario, Gianna Vivaldo, Alessandro Bessi, Fabiana Zollo, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2016. Echo Chambers: Emotional Contagion and Group Polarization on Facebook. *Scientific Reports* 6, (December 2016), 37825. DOI:https://doi.org/10.1038/srep37825

[13] Michael A. DeVito, Darren Gergle, and Jeremy Birnholtz. 2017. " Algorithms ruin everything" # RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, 3163–3174.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[15] Amandeep Dhir, Yossiri Yossatorn, Puneet Kaur, and Sufen Chen. 2018. Online social media fatigue and psychological wellbeing—A study of compulsive use, fear of missing out, fatigue, anxiety and depression. *International Journal of Information Management* 40, (2018), 141–152.

[16] Mark Diaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing Age-Related Bias in Sentiment Analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), Association for Computing Machinery, New York, NY, USA, 1–14. DOI:https://doi.org/10.1145/3173574.3173986

[17] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. 2017. " Be Careful; Things Can Be Worse than They Appear": Understanding Biased Algorithms and Users' Behavior Around Them in Rating Platforms. In *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM)*, Montréal, Canada, 62–71.

[18] Rui Fan, Jichang Zhao, Yan Chen, and Ke Xu. 2014. Anger is more influential than joy: Sentiment correlation in Weibo. *PloS one* 9, 10 (2014), e110184.

[19] Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)* 51, 4 (2018), 1–30.

[20] Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '16), Association for Computing Machinery, San Francisco, California, USA, 2125–2126. DOI:https://doi.org/10.1145/2939672.2945386

[21] Silas Hsu, Kristen Vaccaro, Yin Yue, Aimee Rickman, and Karrie Karahalios. 2020. Awareness, Navigation, and Use of Feed Control Settings Online. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13.

[22] Gawesh Jawaheer, Peter Weller, and Patty Kostkova. 2014. Modeling user preferences in recommender systems: A classification framework for explicit and implicit user feedback. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 4, 2 (2014), 1–26.

[23] John T. Jost, Pablo Barberá, Richard Bonneau, Melanie Langer, Megan Metzger, Jonathan Nagler, Joanna Sterling, and Joshua A. Tucker. 2018. How social media facilitates political protest: Information, motivation, and social networks. *Political psychology* 39, (2018), 85–118.

[24] Haewoon Kwak, Jeremy Blackburn, and Seungyeop Han. 2015. Exploring Cyberbullying and Other Toxic Behavior in Team Competition Online Games. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15), ACM, New York, NY, USA, 3739–3748. DOI:https://doi.org/10.1145/2702123.2702529

[25] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15.

[26] Q. Vera Liao, Moninder Singh, Yunfeng Zhang, and Rachel KE Bellamy. 2020. Introduction to Explainable AI. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–4.

[27] L. I. Lifang, WANG Zhiqiang, Qingpeng ZHANG, and W. E. N. Hong. 2020. Effect of anger, anxiety, and sadness on the propagation scale of social media posts after natural disasters. *Information Processing & Management* 57, 6 (2020), 102313.

[28] Eugene Ndiaye, Tam Le, Olivier Fercoq, Joseph Salmon, and Ichiro Takeuchi. 2019. Safe grid search with optimal complexity. In *International Conference on Machine Learning*, 4771–4780.

[29] William S. Noble. 2006. What is a support vector machine? *Nature biotechnology* 24, 12 (2006), 1565–1567.

[30] Douglas W. Oard and Jinmook Kim. 1998. Implicit feedback for recommender systems. In *Proceedings of the AAAI workshop on recommender systems*, WoUongong.

[31] Michael J. Paul and Mark Dredze. 2014. Discovering health topics in social media using topic models. *PloS one* 9, 8 (2014), e103408.

[32] Brian A. Primack, Ariel Shensa, César G. Escobar-Viera, Erica L. Barrett, Jaime E. Sidani, Jason B. Colditz, and A. Everette James. 2017. Use of multiple social media platforms and symptoms of depression and anxiety: A nationally-representative study among US young adults. *Computers in human behavior* 69, (2017), 1–9.

[33] Aditya Kumar Purohit, Louis Barclay, and Adrian Holzer. 2020. Designing for Digital Detox: Making Social Media Less Addictive with Digital Nudges. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–9.

[34] Anette Rantanen, Joni Salminen, Filip Ginter, and Bernard J. Jansen. 2019. Classifying online corporate reputation with machine learning: a study in the banking domain. *Internet Research* 30, 1 (January 2019). DOI:https://doi.org/10.1108/INTR-07-2018-0318

[35] Joni Salminen, Hind Almerekhi, Milica Milenković, Soon-gyo Jung, Jisun An, Haewoon Kwak, and Bernard J. Jansen. 2018. Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media. In *Proceedings of The International AAAI Conference on Web and Social Media (ICWSM 2018)*, San Francisco, California, USA. Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17885

[36] Joni Salminen, Soon-gyo Jung, Shammur Chowdhury, and Bernard J. Jansen. 2020. Analyzing Demographic Bias in Artificially Generated Facial Pictures. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems Extended Abstracts*

(CHI '20), Association for Computing Machinery, Honolulu, HI, USA, 1–8. DOI:https://doi.org/10.1145/3334480.3382791

[37] Joni Salminen, Fabio Veronesi, Hind Almerekhi, Soon-gyo Jung, and Bernard J Jansen. 2018. Online Hate Interpretation Varies by Country, But More by Individual: A Statistical Analysis Using Crowdsourced Ratings. In *Proceedings of The Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS - 2018)*, IEEE, Valencia, Spain.

[38] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[39] Edwin D. Simpson, Matteo Venanzi, Steven Reece, Pushmeet Kohli, John Guiver, Stephen J. Roberts, and Nicholas R. Jennings. 2015. Language understanding in the wild: Combining crowdsourcing and machine learning. In *Proceedings of the 24th international conference on world wide web*, 992–1002.

[40] Kanokporn Sriwilai and Peerayuth Charoensukmongkol. 2016. Face it, don't Facebook it: impacts of social media addiction on mindfulness, coping strategies and the consequence on emotional exhaustion. *Stress and Health* 32, 4 (2016), 427–434.

[41] Enrico Steiger, Rene Westerholt, and Alexander Zipf. 2016. Research on social media feeds–A GIScience perspective. *European Handbook of Crowdsourced Geographic Information* (2016), 237.

[42] Haridimos Tsoukas. 1997. The tyranny of light: The temptations and the paradoxes of the information society. *Futures* 29, 9 (1997), 827–843.

[43] Selen Türkay, Jessica Formosa, Sonam Adinolf, Robert Cuthbert, and Roger Altizer. 2020. See No Evil, Hear No Evil, Speak No Evil: How Collegiate Players Define, Experience and Cope with Toxicity. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13.

[44] Kristen Vaccaro, Dylan Huang, Motahhare Eslami, Christian Sandvig, Kevin Hamilton, and Karrie Karahalios. 2018. The illusion of control: Placebo effects of control settings. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13.

[45] Anna Vannucci, Kaitlin M. Flannery, and Christine McCauley Ohannessian. 2017. Social media use and anxiety in emerging adults. *Journal of affective disorders* 207, (2017), 163–166.

[46] Bertie Vidgen and Leon Derczynski. 2020. Directions in Abusive Language Training Data: Garbage In, Garbage Out. *arXiv preprint arXiv:2004.01670* (2020).

[47] Vasillis Vlachokyriakos, Clara Crivellaro, Christopher A. Le Dantec, Eric Gordon, Pete Wright, and Patrick Olivier. 2016. Digital civics: Citizen empowerment with and through technology. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*, 1096–1099.

[48] Pamela Wisniewski, Haiyan Jia, Na Wang, Saijing Zheng, Heng Xu, Mary Beth Rosson, and John M. Carroll. 2015. Resilience mitigates the negative effects of adolescent internet addiction and online risk exposure. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 4029–4038.

[49] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, and Morgan Funtowicz. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv* (2019), arXiv-1910.

[50] Heather Cleland Woods and Holly Scott. 2016. # Sleepyteens: Social media use in adolescence is associated with poor sleep quality, anxiety, depression and low self-esteem. *Journal of adolescence* 51, (2016), 41–49.

[51] Melita Zajc. 2015. The social media dispositive and monetization of user-generated content. *The Information Society* 31, 1 (2015), 61–67.