Improving Users’ Algorithmic Understandability and Trust in Content Moderation

Jibon Naher, School of Computing, KAIST, jibon@kaist.ac.kr
Taehyeon An, School of Computing, KAIST, taeohy@kaist.ac.kr
Nitesh Goyal, Jigsaw, teshg@google.com
Juho Kim, School of Computing, KAIST, juhokim@kaist.ac.kr

ABSTRACT

Machine learning (ML) algorithms are actively used for content moderation in many online discussion platforms to keep up with the large volume of content generated by users everyday. However, most of the time, the existence of algorithms in moderation is opaque to the user. Even when the user knows about the algorithm, it sometimes makes wrong or biased decisions, which makes the user feel unfair and dissatisfied, and reduces their trust on both the algorithm and the moderation process. While improving transparency about the moderation decision could be a solution to address this issue, it can not be fulfilled without improving the algorithmic transparency used in the moderation process. However, the complex nature of the algorithm makes it difficult to design for algorithmic transparency, especially when users do not have any background knowledge on ML algorithm. In this position paper, we briefly discuss the risks, challenges, and design considerations for improving users’ understandability and trust regarding the ML output in content moderation. We also present a discussion interface prototype designed to improve users' understandability and trust on the algorithm by improving moderation transparency and by providing the option of exploring and interacting with the algorithm as people write posts.

Introduction

Most online platforms prohibit obviously racist, homophobic, and hateful content. Still, the existence of abusive content is common across online platforms [1, 2]. To reduce potential damage caused by bad actors, different platforms adopt different techniques to moderate their content [3]. These techniques take two primary forms: human moderation and human moderation augmented by automated techniques. In the former case, teams of human moderators including potentially externally contracted workers, and/or a small number of selected users from the platform, manually go through the posts, and remove content that violates the terms and conditions of the platform [4]. Users can also contribute in content moderation via voting or reporting mechanism. However, task load for human moderators is not scalable, and the constant exposure to disturbing content negatively and substantially affects the mental health of moderators [5].

To keep up with the immense volume of content created by users, online social platforms—like Facebook [1], YouTube [2], and Twitter [3]—are known to train and apply machine learning algorithms by compiling large datasets of past moderation decisions on the platform. However, deploying these algorithms without any human oversight can sometimes be problematic; for example, in 2018, Tumblr launched a new anti-porn algorithm to flag pornography, but it was accused of creating chaos by flagging random, nonsexual posts [4]. Nonetheless, machine learning approaches can be especially helpful in saving time and effort of human moderators by algorithmically triaging comments to review. From April, 2019 to June, 2019, 99.3% of comments on Youtube were removed after flagging from the automatic detection [2]. The number of reports on Twitter had decreased from 868,349 in January, 2018 to approximately 504,259 in June, 2018 after it introduced technology to proactively identify offensive content [3]. The New York Times (NYT) recently started using an ML tool to prioritize comments for moderation, and sometimes, approve them automatically. It gives the NYT the opportunity to open commenting for articles, which they had closed before due to the large number of toxic comments [13].

However, algorithms are usually housed in black-boxes that limit users' understanding of how an algorithmic decision is made. The limited algorithmic transparency can cause users' dissatisfaction, lower users' trust in the system, and sometimes lead the user to stop using the platform [6, 8]. Despite the enormous use of ML in content moderation, little is known about how end-users interpret the output of the ML algorithm and how we can help users in understanding the algorithm. As end-users are the central actors in online social systems, from the trust and transparency perspective of the user, it is important to design for reducing the opacity of the algorithm used in the moderation decision making [6, 7, 8, 11]. However, there is also the concern of exposing algorithm process to the user from the fear of malicious uses by bad actors [9, 10].

In this position paper, we argue to design for improving users' understandability and trust on the ML algorithm used in content moderation by improving the transparency of the moderation process, and propose an interface design to serve this purpose. First, we discuss the risks and challenges in using ML algorithms in content moderation. Through this discussion, we identify design considerations for improving users' understanding and trust on the algorithm. Then, we present a web based discussion interface which gives the user the option of exploring and interacting with the algorithm, and the option of providing feedback when the ML output does not match the user's expectation. We discuss how different parts of our design could affect users' understandability and trust on the algorithm.

[1] https://newsroom.fb.com/news/2018/04/comprehensive-community-standards/
[2] https://transparencyreport.google.com/youtube-policy/removals
[3] https://blog.twitter.com/official/en_us/topics/company/2018/evolving-our-twitter-transparency-report.html
[4] https://www.buzzfeednews.com/article/krishrach/tumblr-porn-algorithm-ban

Users' understandability and trust on ML algorithm in content moderation

In this section, we briefly discuss the risks and challenges in using ML algorithms in content moderation from the perspective of users' understandability and trust. Through this discussion, we propose three design considerations for improving algorithmic transparency to the user with the purpose of improving users' understandability and trust.

Risks: *Wrong* or *deceptive output* is a major risk in using ML algorithms for decision making, which can lead to user *dissatisfaction***.** One example of wrong decisions is Tumblr's anti-porn algorithm mentioned earlier. Another algorithm which has caused great controversy and dissatisfaction among users due to its opacity is the Yelp review filtering algorithm. Nearly 700 reports have been filed, mostly from small-scale entrepreneurs who are especially vulnerable to online reviews, accusing Yelp of manipulating its review filtering algorithm to force businesses to pay for advertising in exchange for better ratings [5]. Not only is the algorithm opaque, but the users did not know the existence of the algorithm [10]. When users discover this opacity, it can lead them to suspect that the algorithm is biased [6].

Challenges: Without expertise in AI/ML, it would be difficult for people to interpret and understand the results of ML algorithms [7]. The subjective nature of content moderation decisions, in which people can have different views, makes it more challenging to design to improve users' understandability, trust, and transparency on the ML algorithm used in content moderation.

Design considerations: From the above discussion about risks and challenges of using ML algorithm, we identify three design considerations to improve users' understandability and trust on the ML algorithm in content moderation.

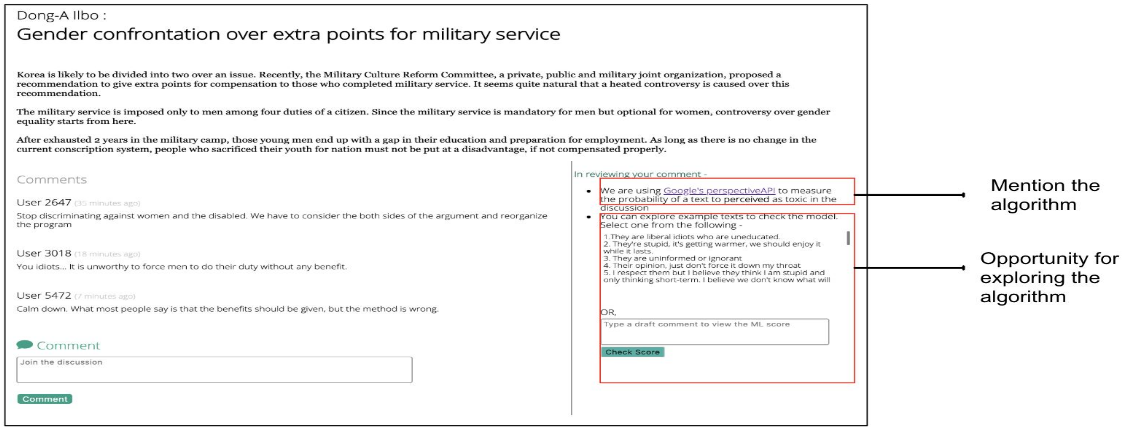
1. *Reduce Opacity in Algorithm Existence*: There should be explicit mentioning about the existence of the ML algorithm with options for exploring the algorithm's performance, which can help users build an accurate mental model.
2. *Reduce Opacity in Algorithm Operation*: Transparency is needed to improve users' trust in the system. Explaining the position and role of the algorithm in the moderation pipeline can be one way of improving transparency.
3. *Collect Users' Feedback*: Users lose trust on the system when they find wrong or biased decisions from the algorithm. Providing a feedback channel where users can file an issue or express their concern can help maintain and improve users' trust. It can also help to reduce users' dissatisfaction on wrong or biased output, and this feedback can be used in updating the algorithm.

[5] https://www.muckrock.com/news/archives/2013/jan/23/businesses-yelp-thug-of-the-internet/
[6] https://www.forbes.com/sites/jimhandy/2012/08/16/think-yelp-is-unbiased-think-again/#7f2251af11d1
[7] https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731

Overview of the Proposed Design:

In this section, we present a web based discussion interface, designed to address the three design considerations mentioned in the previous section. We implemented a discussion interface having a main post, several comments on that post, a comment input box, and with extra design items (e.g., option for exploring the ML algorithm, collecting user feedback) to serve the purpose of our design considerations. In the current design, we use Google's PerspectiveAPI [8], the ML algorithm to score the toxicity of a comment in a discussion, as a content moderation algorithm. PerspectiveAPI is a free tool that uses a machine learning model trained on human-generated comments, labeled as toxic or not by human annotators. This ML algorithm has been used and found effective in the content moderation in several platform recently, e.g., NYT [13], El País [9]. The detailed description of the proposed design is given below.

- **Designing for reducing the opacity in algorithm existence:** In our discussion interface, we divide the comment section into two parts. The left panel has the standard discussion contents (display of comments made by other users, an input textbox for adding a comment, etc.). The right panel includes information to reduce the opacity in algorithmic existence. Although there is no standard for how much information to provide to the user about the algorithm, but providing too much information can have a negative effect on users' trust [12]. In this design, we provide two types of information regarding the algorithm. *First*, it mentions the name and link to a page describing the algorithm in detail, which is Googles' information page for the ML algorithm we are using. *Second*, to help the user understand the output from the algorithm, it gives the user an option to explore the algorithm, by either checking the output on example texts from the dataset used to train the ML model, or by writing their own comments. The latter helps the user in understanding the algorithmic output by changing parts of the text as they want. Figure 1 below shows the discussion interface with this design.



Dong-A Ilbo :

Gender confrontation over extra points for military service

Korea is likely to be divided into two over an issue. Recently, the Military Culture Reform Committee, a private, public and military joint organization, proposed a recommendation to give extra points for compensation to those who completed military service. It seems quite natural that a heated controversy is caused over this recommendation.

The military service is imposed only to men among four duties of a citizen. Since the military service is mandatory for men but optional for women, controversy over gender equality starts from here.

After exhausted 2 years in the military camp, those young men end up with a gap in their education and preparation for employment. As long as there is no change in the current conscription system, people who sacrificed their youth for nation must not be put at a disadvantage, if not compensated properly.

Comments

User 2647 (35 minutes ago)

Stop discriminating against women and the disabled. We have to consider the both sides of the argument and reorganize the program

User 3018 (18 minutes ago)

You idiots... It is unworthy to force men to do their duty without any benefit.

User 5472 (7 minutes ago)

Calm down. What most people say is that the benefits should be given, but the method is wrong.

In reviewing your comment -

- We are using Google's perspectiveAPI to measure the probability of a text to perceived as toxic in the discussion
- You can explore example texts to check the model. Select one from the following -

1.They are liberal idiots who are uneducated.
2. They're stupid, it's getting warmer, we should enjoy it while it lasts.
3. They are uninformed or ignorant
4. Their opinion, just don't force it down my throat
5. I respect them but I believe they think I am stupid and only thinking short-term. I believe we don't know what will

OR,

Mention the algorithm

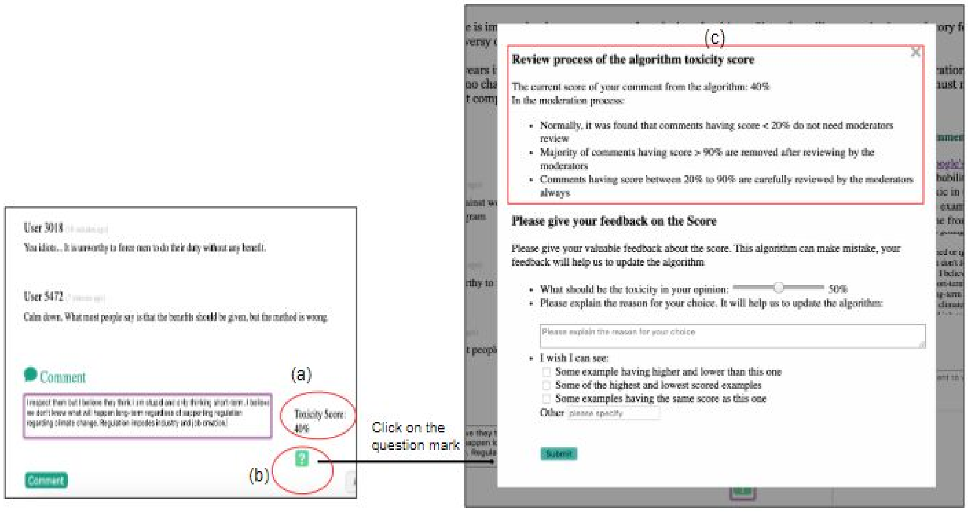
Opportunity for exploring the algorithm

Figure 1: Reducing opacity in algorithm by providing the information about the algorithm used in content moderation process, in the right sidebar of the discussion page.

[8] https://www.perspectiveapi.com/

[9] https://www.blog.google/outreach-initiatives/google-news-initiative/how-el-pais-used-ai-make-their-comments-section-less-toxic/

- **Designing for reducing the opacity in algorithm operation:** To design for improving transparency regarding the position of algorithm in the moderation pipeline, we provide *real-time output* from the algorithm to the user including information about how this output may have an effect on the comment reviewing process. In this design, when a user is typing the comment, it shows the algorithmic output in real-time with a question mark. The purpose of the question mark is to pique users' curiosity and provide detailed information when clicking on the question mark. When the user clicks on the question mark, a modal window opens on top of the main window with information regarding the use of the ML output in the moderation process. Figure 2 shows this design. There are three main parts (a, b, c) in this design, which are marked in the figure.
    - (a): Showing the real-time toxicity score of the typed comment calculated by the algorithm
    - (b): Question mark below the toxicity score to trigger users' curiosity
    - (c): Explanation of the usage of the score in the moderation review process.



User 3018

You idiots... It is unworthy to force men to do their duty without any benefit.

User 5472

Calm down. What most people say is that the benefits should be given, but the method is wrong.

I respect them but I believe they think I am stupid and only thinking short-term. I believe we don't know what will happen long-term regardless of supporting regulation regarding climate change. Regulation impedes industry and job creation.

Toxicity Score: 40%

Review process of the algorithm toxicity score

The current score of your comment from the algorithm: 40%
In the moderation process:

- Normally, it was found that comments having score < 20% do not need moderators review
- Majority of comments having score > 90% are removed after reviewing by the moderators
- Comments having score between 20% to 90% are carefully reviewed by the moderators always

Please give your feedback on the Score

Please give your valuable feedback about the score. This algorithm can make mistake, your feedback will help us to update the algorithm

- What should be the toxicity in your opinion: 50%
- Please explain the reason for your choice. It will help us to update the algorithm:

Please explain the reason for your choice

- I wish I can see:
  □ Some example having higher and lower than this one
  □ Some of the highest and lowest scored examples
  □ Some examples having the same score as this one
  Other please specify

Figure 2: Improving transparency in algorithm operation by providing the real-time algorithmic output and information regarding the moderation process involving the algorithm.

- **Designing for collecting users' feedback:** To reduce users' dissatisfaction in case of wrong or biased output from the algorithm, and increase users' trust in the system, the user has the option of providing their feedback on the output from the algorithm (the modal window in Figure 2, below (c)). In this design, the user can provide indicate the (correct) expected output and provide reasoning behind the expected output. The third question asks the user about what kind of information can help them in understanding the output (e.g., comments having very high or low score, comments having similar score as their typed comment). This feedback can help to improve the design for users' understanding of the algorithmic output, which is included in the first design consideration: reduce opacity in algorithm existence.

Conclusion and Future Plan

This position paper proposes an interface design for improving users' algorithmic understandability and trust in content moderation. We discuss the risks and challenges from the users' understandability and trust perspective, when using a machine learning algorithm in the moderation pipeline. We also discussed design considerations and proposed an interface design based on these design considerations. We plan on concretize the design by doing rapid and iterative prototype designing. We aim to update the design for improving users' understandability by collecting feedback from end-users, and modify the design by iterative UI designing process.

Acknowledgement

This work was supported by the Office of Naval Research (ONR: N00014-18-1-2834), and the Next-Generation Information Computing Development Program through the National Research Foundation of Korea, funded by the Ministry of Science and ICT (NRF-2017M3C4A7065960).

References

[1] Chandrasekharan, Eshwar, et al. "The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales." *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (2018): 32.

[2] Nitasha Tiku and Casey Newton. February 4, 2015. Twitter CEO:"We suck at dealing with abuse." The Verge (February 4, 2015).

[3] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating behavior in online communities. Building Successful Online Communities: Evidence-Based Social Design. MIT Press, Cambridge, MA (2012), 125–178.

[4] Roberts, Sarah T. "Commercial content moderation: Digital laborers' dirty work." (2016).

[5] Sarah T Roberts. 2014. Behind the screen: The hidden digital labor of commercial content moderation. Ph.D. Dissertation. University of Illinois at Urbana-Champaign.

[6] Diakopoulos, Nicholas. "Algorithmic accountability reporting: On the investigation of black boxes." (2014).

[7] Herlocker, Jonathan L., Joseph A. Konstan, and John Riedl. "Explaining collaborative filtering recommendations." *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM, 2000.

[8] Seaver, Nick. "Knowing algorithms." (2014): 1441587647177.

[9] Eslami, Motahhare, et al. "First i like it, then i hide it: Folk theories of social feeds." *Proceedings of the 2016 cHI conference on human factors in computing systems*. ACM, 2016.

[10] Eslami, Motahhare, et al. "User Attitudes towards Algorithmic Opacity and Transparency in Online Reviewing Platforms." *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 2019.

[11] Nitesh Goyal and Susan R. Fussell. 2016. Effects of Sensemaking Translucence on Distributed Collaborative Analysis. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16). ACM, New York, NY, USA, 288-302. DOI: https://doi.org/10.1145/2818048.2820071

[12] Kizilcec, René F. "How much information?: Effects of transparency on trust in an algorithmic interface." *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2016.

[13] Bassey Etim, 2017, https://www.nytimes.com/2017/06/13/insider/have-a-comment-leave-a-comment.html