Chromatify

A Design Brief

Prepared by:
Ambyr Braxton, André Dvořák, Madison Trepanier, Mahmood Jasim, Michael Orlando

Overview of the Issue

Communicating online in the form of comments is not a new trend. It's something that has been occurring since the first online forums were created. These comment sections continued to involve as social media platforms started to grow. Comment sections and threads are a space for users to engage with the original thoughts and posts; to create a more engaged and participatory online community. What once started as a way for people to connect with each other via comments quickly became a space for open toxicity.

From YouTube comment sections to Twitter threads, users who post content to social media are often subject to abusive or toxic commentary from their disgruntled counterparts. A Pew Research[1] study found that about 40% of internet users have been victims of online harassment, specifically in the comments section. Social media platforms are, by far, the most common digital spaces (75% of all abuse) to host these negative interactions for users. This dynamic not only affects personal accounts, but the social media managers of brands, products, and other institutions, many of which become the recipients of simultaneously unwarranted and harmful criticism.

Though many users attempt to ignore such comments,[2] the resulting outcome of these interactions is often destructive for both the recipient and commenter. Research and testimony indicates that negative interactions unfolding on social media are generally associated with negative mental health symptoms.[3] Not only do negative comments and disgruntled users affect the mental health of those reading the comments, they also affect perceived credibility. Studies have shown that negative comments can affect how a reader perceives the trustworthiness or persuasiveness of an article.[4] Businesses also must adapt to shield their employees from potential discriminatory or destructive users.[5]

Meanwhile, angered users put themselves at risk of punishment from the social media platforms they use to share their negative opinions. Without alternate forms of expression built into the design of these platforms, many users who find themselves angered in the moment choose to proceed with

---

[1] Vogels, Emily A. 2021. "The State of Online Harassment." *Pew Research Center*, January 13, 2021. https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/.

[2] Sunderland, Mitchell. 2016. "How Celebrity Social Media Managers Handle the Alt-Right Abusing Their Clients." *Vice News*, October 1, 2016. https://www.vice.com/en/article/8x4agp/how-celebrity-social-media-managers-handle-the-alt-right-abusing-their-clients.

[3] Primack, Brian, et al. 2018. "The association between valence of social media experiences and depressive symptoms." *Anxiety & Depression Association of America*, (June), 784-794. https://doi.org/10.1002/da.22779.

[4] Bogomilova, Aleksandra. 2016. "How reading online comments affects us." *Social Media Psychology*, October 5, 2016. https://socialmediapsychology.eu/2016/10/05/onlineandsocialmediacomments/.

[5] Ho, Pang-Chieh. 2019. "'I've Seen The Worst In Humanity': The Hard Reality Of Being A Brand Social Media Manager." *Digg*, August 14, 2019. https://digg.com/2019/what-being-a-brand-social-media-manager-is-like.

expressing their toxic commentary. Others, still, will choose to harass users because they enjoy the response it garners from those affected.

Who does it impact?

Online comments are crucial to the internet experience. When shopping, users tend to read reviews to help influence their purchasing decisions. Other times users will comment what they think of an article. Brands, restaurants, and online businesses rely heavily on input from their users' experiences; the downside may be that not every review or comment will be positive. The issue of negative online comments mainly involves those who leave negative remarks, and those on the receiving end. In an attempt to create an open space online, without limiting freedom of expression, we want to create a way for people to express their negativity in a way that does not harm anyone else in the process.

Initially, our product was aimed at changing the behavior of "trolls", those who harass others online for the enjoyment of doing so. For this subset of individuals posting negative comments, atypical social rewards—in this case, creating a toxic social environment on social media and reaping the benefits of the ensuing mayhem—override the possibility of behavioral change.[6] These users derive pleasure from the negative environment they create, and therefore, there is very little incentive to cease "trolling."

As a result of this mindset, our research revealed that any direct intervention to correct the behavior of these users may have the opposite effect. Emotional reactions from recipients embolden and affirm the troll's mission to create hysteria. Similarly, platform policy changes and technical design choices quickly become new "games" for these users to learn to circumvent and exploit. Therefore, we quickly had to acknowledge that any design aimed at solving the problem of trolling may prove fruitless.

In a 2014 survey conducted by YouGov, 28 percent of Americans admitted to engaging in malicious online activity directed at somebody they didn't even know and of this 28 percent, 12 percent of posters admitted to having crossed the line so far that they have had their comment removed by a moderator.[7] So why do trolls feel the need to engage in this type of behavior online? We have seen the studies that address the outcomes of being harassed online; psychologists have studied it for years. This research holds significant weight and fuels many of the anti-bullying campaigns we see today. However, what we do not see as often is research behind the why. Why do people troll online to begin with is the source that we believe needs to be discussed in order to create something that has a significant impact on the online community as a whole. A study from Psychology Today took this opportunity to look at

[6] March, Evita. 2017. "How empathy can make or break a troll." *The Conversation*, July 12, 2017. https://theconversation.com/how-empathy-can-make-or-break-a-troll-80680.
[7] Gammon, Jake. 2014. "Over a Quarter of Americans have Made Malicious Online Comments." YouGov. https://today.yougov.com/topics/politics/articles-reports/2014/10/20/over-quarter-americans-admit-malicious-online-comm.

why people troll online. According to their research there is not just one reason, trolling is multi-causal, meaning that it is not caused by any one reason but rather by many reasons that can add up and interact with each other "in a perfect storm to produce trolling."[8] For this reason there is no one cure to change the behavior and depending on the reasoning and severity behind the trolling, they will continue to do so despite intervention because it reeps a response out of the receiver.

At the end of the day, we cannot change trolls for a multitude of reasons. The most significant reason being free speech. As social media continues to advance the idea of where free speech begins and ends is muddy. There have been numerous Supreme Court cases that dealt with free speech and harassment, for example the case of SAXE vs. State College Area School District, however the Supreme Court is now dealing with free speech in the new era of the internet and must make decisions of how free speech applies online. Currently the U.S. Supreme Court is deciding on Mahanoy Area School District v. B.L., case that involves a high school students outburst on snapchat and "whether to issue a sweeping ruling bringing First Amendment law for schools into the social media age or settle for a much more modest decision." [9]

For these reasons we believe that the final iteration of Chromatify is intended for "middle-ground" users involved in sending negative comments. These individuals may be absorbed by the emotion of the moment, may not realize the severity of their action, or at least have the capacity to reflect on their decisions and the impact they have on themselves and others. These users are not always looking to pick a fight; there are many reasons that may spark rebuke from middle-grounders, such as political disputes, combating misinformation, calling out discrimination, and more. Middle-ground users make mistakes, can admit to them, and present a willingness to improve their behavior and online standing. This group of commenters will be the primary audience for our product, offering them an alternative method of expression that allows them to contemplate their negative content, consider the implications of sending it, and offering them the choice to proceed with either approach.

Previous, Current, and Future Attempts

Toxicity in social media space is not a recent phenomenon. As early as 2007, developers and researchers were gradually becoming aware of the futility of the then typical spamming detection algorithms to be effective against identifying toxic commentary in social media posts and messages. What followed was a decade of research in an attempt to address this issue that took a two-pronged approach. Researchers in natural language processing and machine learning investigated and proposed several algorithmic

[8] Stea, Jonathan N. 2020. "Why Do People Troll Online." Psychology Today. https://www.psychologytoday.com/us/blog/writing-integrity/202008/why-do-people-troll-online.
[9] Gerstein, Josh. 2021. "Supreme Court Grapples with Free Speech Case Involving Student's Snapchat Outburst." Politico. https://www.politico.com/news/2021/04/28/supreme-court-snapchat-case-484925.

solutions[10] taking content, sentiment, and context features[11] into account. In conjunction, researchers in social science, social computing, psychology, and human-computer interaction explored different intervention strategies to inhibit users from posting content that might negatively impact others.[12] Such interventions included adding delays[13] when posting toxic commentary to enable users to rethink their actions, informing the hidden consequences of posting such content, and resources to educate[14] potential toxic users to know more about the impact of such posts and refrain themselves from posting. Nextdoor, a social network for neighborhoods, recently implemented an intervention using these ideas. Their system presents the user with a notification when potentially racist content is detected. This notification will give users information as to why this content might be harmful and allows for users to revise their post if they wish. This anti-racism intervention is not the first type of mediation for Nextdoor, in 2019 they introduced an offensive language notification — seeing a 30% reduction in 'incivil content' due to the notification.[15]

One technique explored in the past was disemvoweling, or the act of removing all vowels from a post. The goal here was to create some form of censorship over hate speech, however this does not always mean that users would not be able to understand the post. Boingboing, a site that implemented disemvoweling wrote that, "disemvoweling is a way that BB can say 'you probably don't want to read this, but you can if you try' and also, 'hey, that was out of line, so we will not let it stand as written'."[16] The site implemented this feature as a request you can make to a moderator, or an automatic action if a specific number of users requested it.[16] For example, the quick brown fox jumps over the lazy dog becomes th qck brwn fx jmps vr th lzy dg, or I hate your family becomes ht yr fmly. With a little effort one can interpret these messages, although as the length of the message increases so does the effort required to parse the information. And this minimal abstraction allows for users to quickly ignore the text if they do not want to read potentially hurtful content.

Current social media platforms have a variety of tools to combat toxic behavior, however these tools are typically geared at people looking to filter out content from their own feeds. There is almost a

---

[10] Bretschneider, Uwe, Thomas Wöhner, and Ralf Peters. 2014. "Detecting Online Harassment in Social Networks." *International Conference on Information Systems*, December 2014. https://core.ac.uk/download/pdf/301363213.pdf.
[11] Yin et al. 2009. "Detection of Harassment on Web 2.0." *Content Analysis in the WEB 2.0*, (2009): 1-7. https://www.cse.lehigh.edu/~brian/pubs/2009/CAW2/harassment.pdf.
[12] Dinakear et al. 2012. "Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying." *ACM Transactions of Interactive Intelligent Systems*, September 2012. https://dl.acm.org/doi/abs/10.1145/2362394.2362400.
[13] Van Royan et al. 2017. "'Thinking before posting?' Reducing cyber harassment on social networking sites through a reflective message." *Computers in Human Behavior*, no. 66 (2017): 345-352. https://doi.org/10.1016/j.chb.2016.09.040.
[14] van Laer, Tom. 2014. "The Means to Justify the End: Combating Cyber Harassment in Social Media." *Journal of Business Ethics*, no. 123 (2014): 85-98. https://link.springer.com/article/10.1007/s10551-013-1806-z.
[15] "Nextdoor launches anti-racism notification to prevent discriminatory language." *Nextdoor Blog*, April 19, 2021. https://blog.nextdoor.com/2021/04/19/nextdoor-launches-anti-racism-notification-to-prevent-discriminatory-language/.
[16] "Disemvoweling vs. Flagging." *BoingBoing BBS*, August 2013. https://bbs.boingboing.net/t/disemvoweling-vs-flagging/7360.

universal adoption of some way to unfriend, block, or report users/posts; some platforms have the ability to target your posts to a particular audience; and other platforms have private accounts. For example, Twitter has recently added a setting[17] so that only people mentioned in a tweet can reply. On platforms such as TikTok, Twitter, and YouTube, you can mute posts with specific words, and in TikTok's case you can block a particular sound/song. Facebook has recently formed the Oversight Board[18], a board which makes independent judgments on appeals for content to be restored or taken down. The oversight board is almost like the 'Facebook Supreme Court' as they only come into play when all other options are exhausted. Somewhat similar to the Oversight Board, although incorporating real users into the process, Fan & Zhang[19] have proposed digital juries, a five stage model to govern online spaces.

YouTube has recently tested a new design which does not show the dislike count[20], only displaying this count to the creator of the particular content. YouTube also claims that it is getting better at removing videos which violate its terms of service before users see them. In Q4 2017, 72 out of every 10,000 views were on a video that violated policies, but in Q4 2020 this number is down to 18 out of every 10,000 views.[21] Twitch has begun to take action when harassment is directed at the Twitch community[22], even when that conduct occurs off of Twitch. Some audio platforms like Clubhouse and Twitter Spaces have a different approach as moderating audio[23] only spaces could be more complex. Twitter retains Spaces audio for 30 days, Clubhouse audio is deleted if a session ends without any report, and Discord opts to not collect any audio at all. Outside of the social media platforms themselves, Intel Bleep[24] is an upcoming app that uses AI to detect and filter audio channels, with a transparency angle and rules for users to choose what content they wish to allow — maybe somewhat similar to what gobo.social[25] does for social media feeds.

---

[17] Xie, Suzanne. 2020. "New conversation settings, coming to a Tweet near you." *Twitter Blog*, August 11, 2020. https://blog.twitter.com/en_us/topics/product/2020/new-conversation-settings-coming-to-a-tweet-near-you.html.

[18] "Announcing the First Members of the Oversight Board." *Oversight Board*, May 2020. https://oversightboard.com/news/327923075055291-announcing-the-first-members-of-the-oversight-board/.

[19] Fan, Jenny and Zhang, Amy X. 2020. "Digital Juries: A Civics-Oriented Approach to Platform Governance." *Conference on Human Factors in Computing Systems*, (2020): 1-14. https://dl.acm.org/doi/10.1145/3313831.3376293.

[20] YouTube, "In response to creator feedback around well-being and targeted dislike campaigns, we're testing a few new designs that don't show the public dislike count. If you're part of this small experiment, you might spot one of these designs in the coming weeks (example below!)," Twitter, March 30, 2021, https://twitter.com/YouTube/status/1376942486594150405.

[21] Kastrenakes, Jacob. 2021. "YouTube claims it's getting better at enforcing its own moderation rules." *The Verge*, April 6, 2021. https://www.theverge.com/2021/4/6/22368505/youtube-violative-view-rate-transparency-stat.

[22] "Our Plan for Addressing Severe Off-Service Misconduct." *Twitch Blog*, April 7, 2021. https://blog.twitch.tv/en/2021/04/07/our-plan-for-addressing-severe-off-service-misconduct/?utm_referrer=https://t.co.

[23] Culliford, Elizabeth. 2021. "From Clubhouse to Twitter Spaces, social media grapples with live audio moderation." *Reuters*, February 25, 2021. https://www.reuters.com/article/us-clubhouse-moderation-focus-idINKBN2AP1J2.

[24] Intel Software, "Billions of Gamers Thousands of Needs Millions of Opportunities | GDC 2021 Showcase | Intel Software." Intel Software, March 18, 2021, YouTube video, 43:07. https://www.youtube.com/watch?v=97Qhj299zRM&t=1770s.

[25] "Your social media. Your rules." *Gobo.social*. https://gobo.social/about.

While these approaches help to identify toxic communication and prevent users from posting such contents, the overarching goal of these methods involves protecting the targets of abusive interactions online. In the current form of this project, we seek to explore an alternative approach where users get to post their potentially toxic content online to vent frustration and negative opinions. However, we enable them to convert the toxic content by passing it through an abstraction filter that turns offensive and abusive language into an abstract image. In our next steps section we touch upon the expansion of Chromatify.

The Design Process

As a multidisciplinary group coming from both policy and computer science backgrounds, the Chromatify team came at the issue of online negativity from different perspectives. The two originating ideas came from wanting to address negativity online. The first idea was negativity with respect to abusive comments to certain brands or professional accounts. The other idea was for a way to vent in an online sphere in a pseudo-anonymous and abstract way; screaming into the void without ramifications. We knew that our values and goals reflected the dire need to do something about negativity online. Each of us understands how harmful negativity can be to the person receiving comments. Yet, we also respect the need to have free speech online without being overly censored or moderated, as we are acutely aware of how detrimental content moderation can be for individuals.

These ideas, values, and goals set the stage for what we developed into Chromatify- an intervention that detects profanity and translates it into abstract images, allowing for users to vent their frustrations into the online void, without the negative ramifications that come with receiving potentially harmful comments. How did we get here?

Asking "How Might We" questions was a helpful process in fully understanding what we needed our intervention to do. We asked questions like:

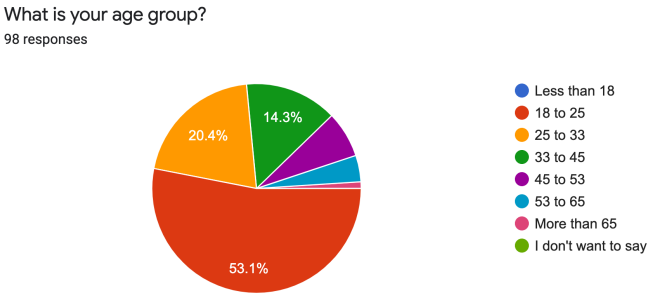
- How might we reduce negativity online? (a broad question)
- How might we keep interactions between users positive?
    - How might we allow brand users to get real time feedback, without negative / harmful commentary that may not be as helpful?
- How might we create an abstract version of negative content?
    - How might we include imagery into text based negative comments?
- How might we appeal to an online community who feels the need to post negative comments?
- How might we assess other ways this tool can be used?

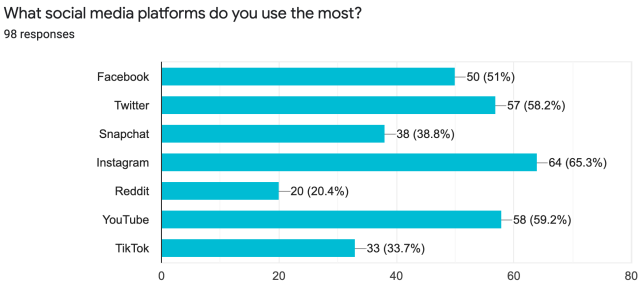
In addition to the “How Might We” process, our team also conducted a survey to assess the scope of potential users. This was a completely anonymous survey due to the nature of the questions, so we did not collect demographic information (aside from age groups). This was an attempt to get feedback from social media users who we envisioned as our main demographic. Our team circulated the survey for a week from April 20th to April 27th, and we received 99 responses. One person indicated that they did not use social media, so we had a total of 98 responses for the survey questions.

We asked questions like

- Do you use social media; what platforms do you use the most?
- Have you ever received a negative comment online?
- Have you ever left a negative comment online?
    - If so what, what was the context; ramifications?
    - If you did not, why?
- Would you use an intervention to allow you to post negative comments without ramifications?

The majority of respondents fell into the age group of 18 to 25. With the 43% of respondents indicating that they use social media 1-3 hours a day. This was followed by 35% using social media 3-5 hours a day.





We also asked respondents to indicate which social media applications they use the most.

Here you can see that the top platforms were Instagram, YouTube, Twitter, and Facebook. With alternative categories including Discord and LinkedIn.

50% of our respondents indicated that they had received a negative comment online, and the majority of cases were in the context of political arguments or when defending their personal beliefs to someone who was questioning them. Some indicated it was upsetting, with others indicating that it had no real impact on them.

We found that the most telling information came from the commentary from respondents when we asked them why they *did not* leave a negative comment online. Some direct quotes are listed below:

- "Worry over backlash/being taken out of context, and also feeling like nothing would be accomplished by arguing over the Internet."
- "My own self conscious & desire to not add more negativity to the world/stop spreading negative troll virus."
- "I was taught 'don't feed the trolls'"
- "Didn't feel like I would change their view or opinion. Fear of cancel culture and mob mentality. Lack of anonymity."
- "Primarily, the understanding that my words and opinions expressed online are difficult to remove completely. Fear of potentially losing employment and/or not earning employment opportunities because of my actions online is a huge reason that generally stops me from posting negative comments."
- "I didn't want to be seen as aggressive by other parents of my kids' friends."
- "I value my job over expressing my opinion about another person. Usually if I see something racist/bigoted I get the urge to call it out, then realize that it's probably pointless and or could be interpreted incorrectly. Fear of being cancelled prevents me from joining any kind of discourse in a negative or accusatory way."

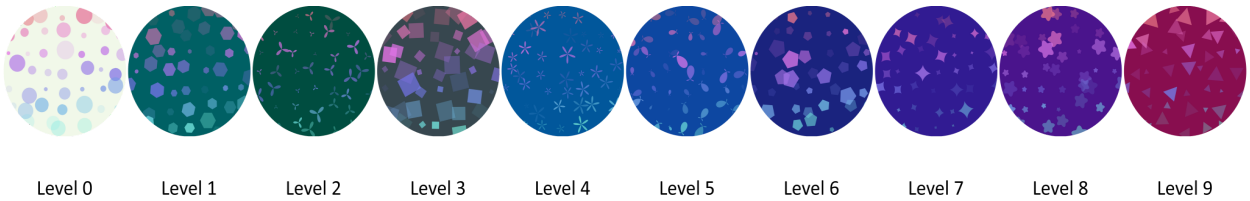
We then asked if people would use an intervention to be able to post negative comments without ramifications. Interestingly enough, 42% said Maybe with Yes and No being tied at 28%. This could be explained by the fact that we did not clearly explain what our intervention was, but this was decided intentionally as we were still shaping what exactly our intervention would be.

After assessing our survey results, we decided that it might not be best to go after the outright "trolls" on the internet. We still like the idea of having our intervention approach content from a user based approach, in which users moderate themselves before posting. These responses indicate that our intervention would be most effective when targeted to the demographic of people who would think twice before posting something negative online, but still might want to vent. We are calling this demographic the "middle ground."

We believe that focusing on those who leave negative comments is more effective than focusing on the targets of these comments. As we previously mentioned, freedom of expression is also key and creating a solution that takes that away is not feasible in the long run. Therefore, our solution is intended for users who are planning to post toxic content, but the benefits extend to both senders and recipients. Middle-ground senders will be able to express their feelings in a way that does not put them at odds

with other users or the rules of social media platforms, while recipients are shielded from potentially harmful content and still have the option to interpret feedback if they choose to.

Chromatify - The intervention

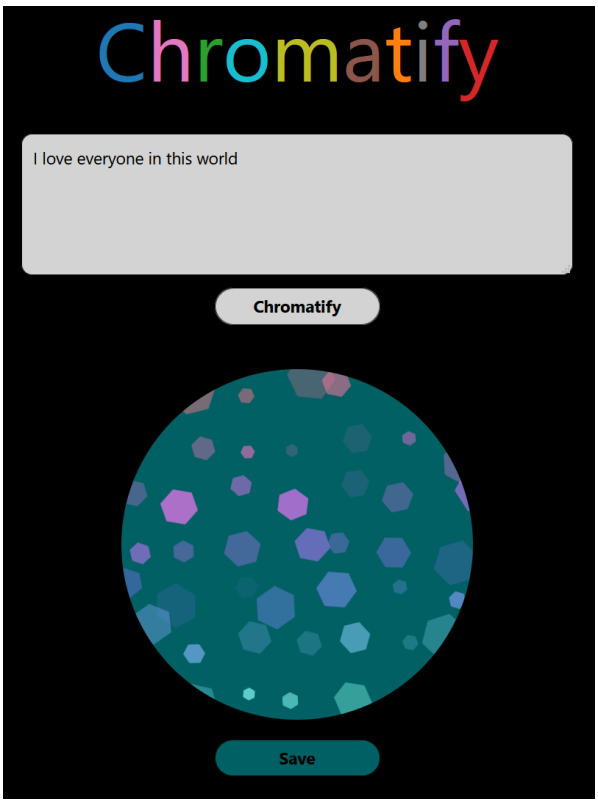


The current version of our intervention, Chromatify, is a website which takes text input, assigns a value between 0 and 9, with 0 being a harmless message and 9 being the most toxic, and generates a corresponding image. Chromatify has a front-end created in HTML, JavaScript, and CSS, and a back-end running a Python Flask application. Any text input in the frontend is sent to the backend where the text is examined to check if there are profanities involved in the text. We used the popular profanity-check library[26] to check the profanity level of the input text. The library uses a linear SVM model trained with two labeled datasets from the hate speech and offensive language dataset[27] and the Kaggle toxic comment classification challenge[28]. The words from these datasets are used to create a bag of words model and upon receiving a piece of input text, the model vectorizes the input string before feeding them to a linear SVM classifier. During the training process, the model learns which words are profane and how profane these words are. These bad words have a higher probability to appear in offensive texts. As such, the training process isolates and learns profane words from all possible words and uses the knowledge to predict future predictions on the input text. This approach is superior to having an arbitrary list of profane words that might not be effective for text input where none of the words in the list match with the words present in the input text. However, the model is only effective for the English language and not languages of other origins. We circumvent this issue in Chromatify by applying a 2500 word denylist where we collected profane words from 15 different languages. We detect if the input text is English, if not, we check with every word from the denylist and if there is a match, we consider such input to be profane. The output of the classifier provides a binary response of whether the input text was profane as well as the probability of the input text being profane from a scale of 0 to 1. We consider each interval of 0.1 to be a discrete probability level where 0.0 means no

[26] vzhou842, "Profanity-check," GitHub Repository, May 23, 2020, https://github.com/vzhou842/profanity-check.
[27] t-davidson, "hate-speech-and-offensive-language," March 29, 2019, https://github.com/t-davidson/hate-speech-and-offensive-language/tree/master/data.
[28] Jigsaw/Conversation AI, "Toxic Comment Classification Challenge," March 20, 2018, https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data.

profanity at all and 1.0 means a heavily profane input text. We use this probability to index a set of shapes and colors based on which the chromatify image pattern is generated in the front-end.

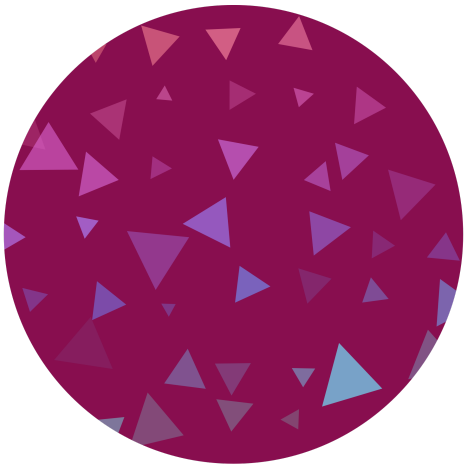
In the front-end, we heavily used the power of CSS to produce different customizable patterns based on user input. To do so, we used the css-doodle package[29] where different patterns can be generated by varying the shapes and colors of the web component. We used a set of 20 different shapes and 10 different colors. We chose basic shapes from the css-doodle package for the 20 shapes. Every profanity level can have two different shapes indexed at random giving users some agency for choosing which shape they would prefer. The shapes are selected in a way where the less profane the words are, the smoother the shape. For example, the profanity level of 0.0 is either represented with a circle or an octagon with relatively smoother edges throughout their circumferences. As the profanity level increases, the shapes become sharper. For example, the profanity level of 1.0 is mapped to either a triangle or a hypocycloid with 3 sharp jagged edges.



The profanity level is also used to map different colors as the background of the patterned image. The color levels for less profane words are selected from softer blue/green hues and as the profanity level rises, the color hue gradually switches to more red. We used Hue-Saturation-Lightness color representation for the colors. We experimented with various colors representations and eventually used the Google Material Design colors[30]. In contrast with the shapes, we used 10 different colors for 10 different profanity levels.

To get a greater sense of the Chromatify user experience, the image on the left shows part of the website, with output for "I love everyone in this world". This is assigned a value of 0 by Chromatify, and the corresponding image is one with soft, non-threatening colors. Users can click on the image to update and generate a new image until they click 'save' to download. To contrast this image, we can take the example of Mahanoy Area School District v. B.L. and generate an image for the text, "Fuck

[29] "<css-doodle />." Accessed April 13, 2021. https://css-doodle.com/.
[30] "Material Design Colors - Material Palette." Accessed April 20, 2021. https://www.materialpalette.com/colors.



school fuck softball fuck cheer fuck everything."[31] Chromatify assigns a value of 9 to this text, and the corresponding image is one with deeper, more dramatic color. The shape is now a sharper, more angular triangle, which also adds to the image's intensity. Again, the focus of this current version of Chromatify is on 'middle-ground' users who want to express their views, but want to mitigate the personal, professional, or societal ramifications of that post. Therefore, in its current form, Chromatify requires the most user participation as users need to go to an external site to generate and download these images, however we touch upon further integration and expansion of the product in the next steps section — Chromatify will only improve.

Design Limitations & Next Steps

There are two major caveats to the inner functionality of Chromatify. Firstly, the classification algorithm used in the system prioritizes speed over accuracy. For example, less common forms of profane words such as, "f4ck you" or "you b1tch" will be difficult for the system to classify due to the lack of occurrences of such variations in the training corpus. At best, the library should be considered to be a heuristic and not unequivocal truth. Second, the lack of trained models proficient in classifying profane words from other languages exacerbate the situation. As such, at this state, the system penalizes profane words from other languages more harshly as the mention of any words from the profane list is immediately considered to be a heavily profane input with the profanity level of 1.0. In future, superior models with superior multilingual text embedding could be used to improve the performance of the profanity detection model, such as BERT or GPT-3. However, we must emphasize that the detection of profane language from various word forms designed to trick the system is not the primary emphasis of Chromatify. Chromatify provides an option for users to masquerade their potentially profane message into aesthetically pleasing patterns and help them avoid unwanted ramifications online while enabling them to express themselves in a comparatively harmless and less evocative manner.

The current Chromatify design also suffers from lack of variations. The rigidity of the css-doodle web component is a limiting factor when it comes to dynamically create responsive patterns suitable for a broader set of design options for users to choose from. The current Chromatify version is designed as a

[31] Howe, Amy. 2021. "Student's Snapchat sets up major ruling on school speech." *SCOTUSblog*, April 27, 2021. https://www.scotusblog.com/2021/04/students-snapchat-sets-up-major-ruling-on-school-speech/.

proof of concept and in the future, we will investigate other libraries and frameworks for enabling more dynamic design ideas with an additional emphasis on user agency so that the users have more control over the patterns generated. These options could include JavaScript canvas or similar platforms to create dynamic shapes and designs.

In terms of the user experience, next steps would be to integrate the service into the user's browser to provide easier access. In the current Chromatify form, the user needs to access an external site to generate their image. A browser extension could provide simplified access to the user by generating these images without leaving the website they are visiting. It also allows us to serve more users, as with any intervention it is important to provide the smallest barrier to entry as possible. After completing this, a following step would be to develop a way for users to apply this filter to an entire website. Sometimes we all need a day (or more) without viewing toxic content on the internet — wouldn't it be nice to be able to filter out and turn certain posts to their Chromatify representation? Users would be able to choose a threshold where any content above that threshold would not be textually displayed, and this would be strictly a local change. After this, adding greater detection for content outside of profanity and in various languages will be the next priority.

Conclusion

The Chromatify intervention is just one possible step in creating an online environment that focuses on positive interactions, rather than negative ones. As a team, we are excited about the potential of other applications related to the intervention we have designed with Chromatify. Freedom of expression online is something we do not want to lose in a new era of moderation, and this is an attempt to encourage users to self-moderate. We are aware of the limitations of this approach, but we are also enthusiastic about the initial prototype. We know there is no use in changing trolls and their behavior, so this is an attempt to reach other users who might be tempted to contribute negatively in online spaces. We envision an online community where venting is allowed, but not harmful and with Chromatify this is possible. Not only can you scream into the void with as many profanities as you want, you also get a pleasing image as a result!