Abstract:

In contemporary society, individuals have gathered on many social media websites in order to express their personal thoughts and opinions to an anonymous public. However, there are many cases in which these thoughts can be framed as attacks or insults, thus garnering controversial backlash. Hate speech is not uncommon in America with its almost polarized citizens. Individuals, especially those belonging to far right-wing groups such as the Alt-Right, will provide hateful remarks on cultural, racially, political, and social topics throughout multiple platforms on the internet. Facebook and Twitter are two of the main social media sites that deal with this issue on a day to day basis.

What has been done is that websites have begun to ban these groups from their sites, forcing Alt-Right groups to abandon mainstream platforms to sites with fewer filters; furthering the marginalization of varying groups online. The consequence is that hate speech from these individuals will disappear from the eyes of the general public, but still continue to exist elsewhere. People want to express their ideologies freely, but with censorship, many feel they are being silenced and treat unfairly. The challenge these tech companies continue to face is the issue of allowing hate speech to go through the cracks or eliminating it completely that it hinders opinions and expression. If they do wish to proceed in eliminating hate speech online, they can counter-attack them by use of hate speech detection and filtering tools. However, even then, there are issues and challenges these companies continue to face that make these elimination methods far from perfect.

With information on the internet in such abundance, sites like Facebook have developed algorithms to filter and generate a selected feed that is suitable for each user based on their search history, 'liked' content and preferences. This method has changed the way each and every individual view and traverses the world wide web through a more personalized system. However, information shared online still has its limitations and restrictions.

Information that is deemed inappropriate according to a host website's standards and ideology is blocked or deleted so that such controversial or offensive material does not spread to other users. Many news sites have created filter bubbles by organizing discourse through the selection, emphasis, and exclusion of one or more aspects, a process known as framing.[1] Framing defines what issue is relevant or not. Also, due to the algorithm creating a personalized feed, users are less likely to be exposed to any objective or subjective ideas or thoughts that are not in

[1] Nguyen, Tien T., Pik-Mai Hui, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan. "Exploring the filter bubble." Proceedings of the 23rd international conference on World wide web - WWW 14, 2014. doi:10.1145/2566486.2568012.

line with their beliefs, unless they search it up for themselves. Forums and social networking sites that align with a certain ideology will appear on the feeds of individuals who share those beliefs and bring them together. As a result, spaces known as Echo Chambers, rooms where only one particular ideology is reflected back and forth, are created. Many find these online spaces more appealing thus have turned away from traditional mainstream media for a more digital environment to engage in.

Technology is constantly being developed and improved, and the focus on digitalization and translating everything to an online space has become more widespread. Thanks to mobile devices such as smartphones and tablets, social networking is more accessible and influential. The convenience of mobile technology has encouraged an increasing number of individuals to use the internet as their main source of information gathering. Even businesses, news networks, and political outlets have seen the benefits of having an online presence, thus they conduct a significant amount of work on the internet in order to strengthen their authority and reach a larger audience. However, this emphasis on digitalization has had major effects on people's perception of reality, which is now based on belonging to a particular community of practice.

Communities of practice refer to groups that distinguish themselves through a social organization in which people as a group learn together. Any individual can belong to one or multiple communities, and members of these communities are brought together through common goals, shared beliefs, and mutual engagement. They view and analyze events and experience from the same perspective. These like-minded interactions have resulted in highly polarized discussions when users of one community engage with other communities, thus resulting in the marginalization of groups. Presently, the strong division between the American left and right wing ideologies can be blamed on these filtering algorithms and social media. The differing conversations on the internet have blurred the lines between a system of ideas that should be protected under the First Amendment and something as dangerous as the spread of hate speech.

Issie Lapowsky, a member at WIRED Magazine, believed that tech companies do not have to be placed under the net of First Amendment, since they are not the government and therefore free to police speech in the manner they deem fittest.[2] This allows them to regulate any information passing through their domains without having to abide by government regulation.

After the 2016 election of President Donald Trump, Americans saw a significant political divide amongst its citizens, as well as increased media attention on the topic of hate speech. President Trump never restrained himself from expressing and sharing his personal, but also

---

[2] Lapowsky, Issie. Tech Companies Have the Tools to Confront White Supremacy. *Wired,* 2017. Accessed from https://www.wired.com/story/charlottesville-social-media-hate-speech-online/

controversial, thoughts on the internet. Many of his tweets on Twitter have varied from racists insults to dangerous threats to other countries. While these responses can be inappropriate and stigmatizing, President Trump himself made comments of this nature excusable. Right-wing individuals have begun to speak more freely of their conservative ideals; feeling a sense of safety against criticisms from a more liberal left-wing population.

Donald Trump's mistrust of traditional "mainstream media" has changed the present media landscape. The allowance of "fake news" or "alternative news" from himself, as well as his own cabinet, has spawned sites such as Breitbart, a news network that provides and spreads conservative and alt-right news, has increased alt-right movement support.[3] Coined by Richard Bertrand Spencer, the term Alt-Right, or the alternative right, is a loosely defined group of people that are made up of far right-wing extremists and white nationalists. These individuals reject the mainstream concepts of political correctness or social justice as these hinder their goals for the preservation of western civilization and white ethnonationalism. While these groups have existed for years, spreading their message in "safe spaces" with websites such as Twitter, Breitbart News, and Reddit, they were not given much consideration by the general public. However, this changed during a far-right rally in Charlottesville, Virginia on August 2017.

The Unite the Right Rally was a scheduled rally in which its goal was to protest against the city's controversial decision to remove America's Civil War Confederate monuments and memorials from public spaces. During this rally, it was to oppose the removal of a statue of Robert E. Lee from Emancipation Park. Much of these protesters included neo-Nazis, Trump supporters, and individuals from the Alt-Right; all of which chanted many racist and antisemitic remarks during the march. The main issue arose when protesters were faced with counter-protesters who were in favor of the removal of these statues. On August 12, a man belonging to a white-supremacist group drove his car into a crowd of counter-protesters resulting in 19 injuries and the death of a 32-year-old woman. In total, the two-day rally resulted in 35 casualties.

Following the incident, many people have claimed what happened in Charlottesville to be an act of domestic terrorism and a hate crime, and opposition towards these groups rose especially on the web. Many websites, as well as service apps like Uber, Airbnb, ApplePay, Discord, GoFundMe, and Paypal,[4]increased their ban on Alt-Right behavior. Alt-Right groups,

---

[3] Cadwalladr, C. "Robert Mercer: The Big Data Billionaire Waging War on Mainstream Media." Accessed November 14, 2017. https://www.theguardian.com/politics/2017/feb/26/robert-mercer-breitbart-war-on-media-steve-bannon-donald-trump-nigel-farage

[4] Avi Selk, "A Running List of Companies that no longer want the Daily Stormer's business," *The Washington Post* (August 16),

feeling their views attacked and unwelcomed more heavily, followed the Daily Stormer's lead and tried banding together to form their own internet

The Daily Stormer was one of the Alt-Right websites that were dismissed by main online platforms. The website had posted an article making derogatory comments about the woman who was killed in Unite the Right rally, which prompted GoDaddy to kick the website off their servers. While it would be easy for the Alt-Right group to hook up a server to the internet, without a domain name, they can't function online. Domain names are regulated by ICANN, a multinational organization that allocates management of generic top-level domains to registries. However, many Alt-Right groups also faced difficulty purchasing domain names as a registry organization due to lack of support from mainstream networks on the basis of insinuating hate speech and promoting violence.

Another significant setback for the Daily Stormer was when Cloudfare, a service that accelerates and secures websites, also booted them out. Currently, the Daily Stormer is receiving Distributed Denial of Service (DDoS) protection from BitMitigate, a small startup company. There have been discussions from site contributors wanting to create their own alternative to Cloudflare as well as other parallel platforms. For now, the Daily Stormer resorted to using the Tor network and now operates on the dark web[5], a space notorious for having no restrictions on controversial topics.

On Gab, a social media website that acts as an alternative to Facebook, users are free to openly share racist remarks or memes. According to Gab's CEO, Andrew Torba, they promote raw, rational, open, and authentic discourse online with no intent on pushing their definition of what is hate speech on users.[6] With popular hashtags on Gab such as #HillarysHealth and #HitlerPickUpLines, it's evident that alt-right groups can speak freely without the fear of being kicked off the site. Hatreon is also a crowdfunding site that doesn't have any limitations on hate speech and has attracted controversial figures such as alt-right and neo-Nazi leaders.[7]

In contrast, Twitter and Facebook have already taken extremely aggressive stances to curb the activities of extremist groups such as ISIS, and the tools used by these companies can be extended to others as well. Instagram developed a filter that automatically deletes specific words

---

https://www.washingtonpost.com/news/the-switch/wp/2017/08/16/how-the-alt-right-got-kicked-offline-after-charlottesville-from-uber-to-google/ (Date of Access: November 2, 2017).

[5] Ibid.

[6] Emma Gray Ellis, "Gab, the Alt-Right's very own Twitter is the Ultimate Filter Bubble," *Wired* (September 14, 2016), https://www.wired.com/2016/09/gab-alt-rights-twitter-ultimate-filter-bubble/ (Date of Access: November 29, 2017).

[7] William Hicks, "Meet Hatreon, The New Favorite Website of the Alt-right," *Newsweek* (August 4, 2017), http://www.newsweek.com/hatreon-alt-right-richard-spencer-andrew-anglin-white-nationalism-white-644546 (Date of Access: November 2, 2017).

from users' feeds; and most interestingly, users can turn this feature on or off.[8] So, users who are more likely to feel offended by certain types of speeches, can simply turn the option on, and filter all the materials that they find offensive.

Many technology companies have also implemented machine learning tools into their AIs in order to detect blacklisted words found on their sites. Companies can use sets of variable-mapping and input them into the machine with desired outputs in mind- and train the machine to perform efficiently and accurately. In other cases, the machine is exposed in a certain environment, and by learning from trial and error can be trained to make specific and accurate decisions. This explains why in many online chat bots, the bot will often ignore or divert the attention from any derogatory topics pertaining to religion, sex, gender, politics, etc. However, there are issues that come with the system and machine learning can easily be manipulated or altered from their intended goals.

An example of such case was the Tay bot in 2016. On March 23, 2016, Microsoft Corporation created and released an Artificial Intelligence chatbot known as 'Tay' onto Twitter. According to the creators, the AI was programmed to mimic the speech of a 19-year-old girl with the goal of becoming smarter by learning from the conversations it encounters from real humans online. [9]Unfortunately, within the 24 hours, the bot was released, Tay had begun to behave inappropriately making racist and sexist remarks, and harassing other twitter users. After 16 hours, the bot was taken down, with Microsoft blaming internet trolls for Tay's behavior. However, it is shown that unlike Microsoft's other AI Cortana, the creators of Tay did not code any appropriate responses to blacklisted words. In turn, the creators unintentionally allowed the AI to become racist. The idea was to make the machine learn to respond like a human from the questions it was asked by humans. Tay did just that. Knowing the bot lacked countermeasures for certain words, this posed a serious design flaw. Tay failed not only because it was exploited by human curiosity on how far artificial intelligence can go, but also because of the programmers' lack of caution when training the machine on how to deal with hateful people online.

A combination of intuitive programming languages and more complex algorithms are needed to help identify trolling or offending patterns on the internet. Subsequently, eliminating this offensive or inflammatory content from websites allowing users to generate content. The Anaconda (Python 3.6) version is a dynamic programming language that has been made popular

---

[8] Zhong, Haoti, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J. Miller, and Cornelia Caragea. "Content-Driven Detection of Cyberbullying on the Instagram Social Network." In *IJCAI*, pp. 3952-3958. 2016.

[9] Sinders, Caroline. "Microsoft's Tay is an Example of Bad Design – caroline sinders – Medium." Medium. March 24, 2016. Accessed November 15, 2017. https://medium.com/@carolinesinders/microsoft-s-tay-is-an-example-of-bad-design-d4e65bb2569f.

because of its high-level functionality and its ability to process natural language data using a Natural Language Toolkit. Using Anaconda, a Troll Identification Algorithm can be created, that is specifically aimed at targeting users and keywords that are misleading, offensive or associated with the nonsensical information.[10] The design of technologies and algorithms is aimed to improve overall identification of inflammatory content, while also allowing the communication technologies to expand without hurting the freedom of expression by over-censoring, and still preventing hate speech to become overbearing.

There have been constant efforts to design abusive or inflammatory speech filters that seamlessly identify sensitive content and filter it out using a token list.[11] However, such methods are outdated and rely on a human generated and populated blacklist for the filtration of content. Instead of this method, which is often deemed inefficient, advanced technologies such as Artificial Intelligence (AI) can be used to detect abusive content in social media. A number of methods that leverage the AI technologies have been identified, and these can be used to automate the data and content filtering process to make it more accurate and efficient. These include dataset balancing, which boosts detection of abusive content on social media; feature reduction, which aims at trimming the features of programs with large feature sets, to improve efficiency while still maintaining detection accuracy; use of generic structural features that can be used universally across datasets.[12]

Hate speech detection has become a helpful, but challenging tool in today's media-based society. The filtering of controversial statements and words have evidently pushed left-wing ideology in the forefront, neglecting right-wing ideology. Thus groups like the Alt-Right have abandoned mainstream platforms to sites with fewer filters, furthering the marginalization of the online population. People want to express their ideologies freely as the first amendment states, even if it's considered hateful, and with censorship, many feel they are being silenced by the majority. However, if their controversial ideologies and actions remain the same, social platform providers have no choice but to keep the ban in place. In turn, the alt-right groups will continue distancing themselves from the liberal stage and vice versa. A possible solution in the elimination of hate speech from these groups is the continued reliance on machine learning and

---

[10] Tayade, Pooja M., Shafi S. Shaikh, and S. N. Deshmukh. "To Discover Trolling Patterns in Social Media:
Troll Filter." (2017).

[11] Snyder, Peter, Periwinkle Doerfler, Chris Kanich, and Damon McCoy. "Fifteen Minutes of Unwanted Fame: Detecting and Characterizing Doxing." *strategies* 10 (2017): 13.

[12] Chen, Hao, Susan Mckeever, and Sarah Jane Delany. "Harnessing the Power of Text Mining for the Detection of Abusive Content in Social Media." In *Advances in Computational Intelligence Systems*, pp. 187-205. Springer International Publishing, 2017.

AI technology, as it has the ability to monitor vast amounts of data over multiple platforms. The consequence is that hate speech will eventually disappear from mainstream sites, but instead cultivate in spaces with fewer restrictions.