The Efficacy of Training Data in Social Media Post-Flagging Algorithms
(Technical Project)

The Propagation of Hate Speech and the Implications of its Moderation
(STS Project)

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Madeleine Ashby

Spring, 2022

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Kent Wayland, Department of Engineering and Society
Rosanne Vrugtman, Department of Computer Science

General Research Problem

The unconscious mind, although entirely composed of naturally occurring processes unavailable to introspection, plays a remarkable role in affecting behavior and emotions (Ricee, 2021). As such, the brain is capable of making very impulsive decisions without conscious awareness or understanding. Most of the time, these blind decisions are simply the brain carrying out daily motor functions, but this is not always the case. The brain learns from what it is presented with, even encoding information into the unconscious that has not passed logical validation in conscious thought. In fact, it is possible that the unconscious is a mind of its own, thinking before the conscious mind is able to form its own objective response. Consequently, your mind makes subconscious, illogical, yet quick determinations based on whatever information you are exposed to, regardless of that information's veracity. This can result in impulsive thoughts and behaviors as reactions to those stimuli– useful for fast thinking, but at the risk of incorrectly making assumptions without even knowing. For instance, if you are raised in a neighborhood in which all families are of the same culture, you would likely be more receptive to familiar-sounding names than to those of other cultures ("Implicit Bias", 2020). This pre-disposition leading to a cognitive "shortcut" is called implicit bias.

While culture and society as a whole would fail to progress without our ability to make quick judgements, humans are increasingly susceptible to implicit bias as a result. It is our tendency to seek out patterns and simplify the world that plays a significant role in shaping implicit biases (Cherry, 2020). Unacknowledged and undisputed, this bias can lead to severe discriminatory actions in the workplace, in the media, and in daily life.

One of the most prominent manifestations of implicit bias is hate speech: where groups of people are discriminated against and targeted with vicious stereotypes and violent ideas. Hate

speech is an explicit projection of bias and prejudice, first conceived when those with implicit biases are radicalized, then resulting in the development of implicit bias in others and eventually, radicalization. Hate speech flows through social media, and although some social media sites implement algorithms to flag and limit hate speech, these imperfect algorithms feed back into the propagation of implicit bias. Consequently, hate speech is the pinnacle of a growing cycle of implicit bias that augments hate and discrimination, all the while contaminating society's view of reality with growing personal biases, only leading to more hate. Implicit bias is the root of hate, and efforts to stop hate speech are being thwarted by the biases of those fighting against it.

Technical Research Problem

The Efficacy of Training Data in Social Media Post-Flagging Algorithms

What aspects of training data are valuable to creating a useful post-flagging algorithm, and how accurate are these algorithms?

With the recent influx of machine learning into everyday technology, it is imperative that we understand how models are constructed in order to properly evaluate their ethical value. In its simplest form, machine learning is a type of artificial intelligence that teaches technology to "think in a similar way to humans: learning and improving from past experiences" (Algorithmia, 2020). The adoption of these machine learning algorithms facilitates the mechanization of processes that were previously only carried out by humans. At a fundamental level, these models are trained to predict outcomes, and are evaluated using several metrics such as accuracy. A heavily present example of a machine learning algorithm we interact with every day would be social media post-flagging algorithms including misinformation detection algorithms or hate speech recognition algorithms. However, neither of these algorithms would exist without one key element: data.

Data used by machine learning classifiers is split into three sets: training, validation, and testing (Buzz Blog Box, 2020). I will focus on training data, as it is the primary data that teaches the algorithms to learn and predict on their own. The creation of an accurate model is dependent on the training data being as unbiased as possible. All data has an opinion, thus rendering all data biased to a certain degree. Alas, we are left with the question: what aspects of training data are necessary to create an unbiased hate-speech classifier?

To answer this question, I will conduct research in several areas, the first being how a machine learning model processes training data and actually learns to make predictions and classifications. This will prompt further research on how changes in training data can affect the way the model learns. For instance, I am particularly interested in how small differences in training data used on the same model can create varying levels of sensitivity. Furthermore, I plan to get a firmer grasp on what makes "good" training data and how we can minimize bias. This research will delve into a few spheres: data availability, prevalence rates in a dataset, and variable types and their compatibility with different models. That being said, I would also like to survey state-of-the-art hate-speech classifiers. I will read scientific journal articles and reports to understand the different types of models compatible with hate-speech recognition, and will evaluate their efficacy using several evaluation metrics such as accuracy, sensitivity, and F1 score. These metrics are particularly relevant to the evaluation of hate-speech classifiers because they are most reflective of false positives and false positives produced by the classifier. Through this comparison, I will discover each model's strengths and applications, ultimately allowing me to draw conclusions regarding the important aspects of training data and the effectiveness of the algorithms they teach. I do believe there may be some constraints on this research, as it could be

difficult to access training data for classifiers due to privacy laws, which may limit my ability to draw conclusions on which aspects of training data are best in creating a hate-speech classifier.

STS Research Problem

The Propagation of Hate Speech and the Implications of its Moderation

How does hate speech on social media lead to radicalization, and can content moderation continue to exist without inhibiting free speech?

Rooted in racism, hate speech surges through society, primarily in the form of social media posts. The most immediate effects of this are obvious – increased racial tensions and a magnified sense of isolation, ultimately jeopardizing the cohesion of society. Unopposed, online hate speech can facilitate the establishment of echo chambers which lead to further polarization and mob mentality. In more extreme cases, communities mobilize and the hate shifts offline, to real life, and people are driven to commit violent acts. The existence of unregulated platforms such as 4chan and Gab serve to amplify these problems, and the murders of several members of minority groups on behalf of neo-Nazi hub *The Daily Stormer* only further exemplify this point (Lavin, 2018).

In recent years, progress towards comfortable, hate-free platforms has been made through the creation of hate speech detection algorithms, but this task hasn't been easy. Widespread disagreements regarding the definition of hate speech serve as an obstacle to creating an algorithm that detects it (MacAvaney et al., 2019). This results in data of varying opinion – the data will be biased based on the definition of hate speech it uses – and makes it difficult to train a model that predicts and acts fairly. Moreover, many automatic detection algorithms neglect societal context in the posts they are reviewing and can, in turn, serve catalytic roles in furthering the systemic racial disparity (Dawson, 2020). For instance, algorithms that do not consider the

context (such as whether the text is dehumanizing or innocuous) in which a group identifier (such as “gay” or “black”) is used often have a higher rate of misclassification than those that do consider context. Furthermore, the algorithms have been subject to scrutiny, as society weighs the democratically imperative value of free speech against the value of safety. This begs the question: how does hate speech on social media lead to radicalization, and can social media content moderation continue to exist without inhibiting free speech?

Background

In recent years, the mobilization of radical hate groups has resulted in several violent acts such as the mass shooting at the Tree of Life Synagogue in 2018. Researchers at the Anti-Defamation League have argued that the tragedy should not have been surprising; perpetrator Robert Bowers had been posting dehumanizing and threatening anti-Semitic messages on social media site Gab for weeks leading up to the event (Guthrie, 2018). Several similar acts have occurred since, including the attack at the ChristChurch Mosque in New Zealand in 2019. According to Dr. Ariel Koch (2020), the high accessibility of platforms similar to Gab is the driving force behind widespread mobilization of extremist groups. In response, both manual and automated hate speech moderation is currently being implemented by several social media companies such as Instagram, Facebook, and Google. Posts reported by users are aggregated and sorted into queues for the moderators to review (Chotiner, 2019). They reference company policies to make a quick, binary decision: the post is offensive and hateful, or it isn’t.

This moderation poses a few substantial issues. First, the hate-speech detection algorithms are biased. Automated moderation utilizes algorithms that often neglect context and are consequently highly sensitive models that misclassify posts on both sides (Dawson, 2020). Secondly, the moderators are biased. Manual moderation utilizes humans who are

inherently biased and become desensitized to the content over time (Chotiner, 2019). Lastly, the process of removing a user's thoughts and opinions from a platform is sometimes viewed as an infringement upon freedom of speech, possibly even further radicalizing those affected by censorship and their followers.

Data Collection and Methods

I will be examining several academic sources including journal articles, magazine articles, and scientific studies. Additionally, to ensure the data is of high quality, I will be using accredited and refereed sources. My research will be framed as an ethical analysis, with the primary argument being analyzed is the belief that the mobilization of extremist groups from online platforms to real-life riots is a threat to public safety and should be moderated. Thus, I will research and identify specific occurrences of online hate speech leading to real-life violence. These examples will be used to illustrate the severity of radicalization of extremist groups and will call attention to the need for understanding this topic.

The primary counterargument to this assertion that will be analyzed is the belief that the moderation of hate speech imposes a limit on free speech (Calvert, 2006). Ben Franklin is quoted as saying, "those who would give up essential Liberty, to purchase a little temporary Safety, deserve neither Liberty nor Safety." To evaluate this position, I will analyze arguments that content moderation is a limitation on the societal value of free speech. Furthermore, I believe a survey of legislation surrounding free speech and limitations imposed on it would strengthen my understanding of this viewpoint. This exploration will seek to understand the growing argument that social media platforms are mostly publicly-owned companies that serve as the only political public forum, so even though they are not governmental agencies, they may need to have moderation policies regulated by the government to preserve the integrity of the

First Amendment. Additionally, I will research company moderation policies that govern the classification of posts as hate speech. This will allow me to ascertain if there are connections between the transparency and rigor of a policy and the ability to distinguish hate speech from innocuous speech (Vaidya et al., 2021).

With all of this knowledge, I plan to evaluate both primary arguments in order to determine whether hate-speech on social media truly causes radicalization, and if content moderation of such hate speech can continue with adversely affecting freedom of speech. My deeper understanding of the matter will hopefully provide enough context and information to allow legislators, social media companies, citizens, and social media users to decide if hate speech moderation is useful and worth the potential costs to combat bias propagation in our society.

Conclusion

This study aims to analyze the debate over whether hate-based violence is a consequence of hate speech on social media and if content moderation can exist without fundamentally altering the way our society values freedom of expression and freedom of speech. This debate is pertinent to present society as implicit bias is spearheading a cycle of bias propagation and hate that ends in violence, and it seems that the current method of inhibition against this cycle— content moderation— is composed of some of the same biased elements which it seeks to eradicate. Content moderation may compromise our societal value of freedom of speech, but it does so in an effort to subdue bias, stall hate, and preserve peace.

Bibliography

Algorithmia. (2020, March 26). The Importance of Machine Learning Data. Algorithmia. https://algorithmia.com/blog/the-importance-of-machine-learning-data

Anti-Defamation League. (n.d.). Gab and 8chan: Home to Terrorist Plots Hiding in Plain Sight. Anti-Defamation League. Retrieved April 2, 2022, from https://www.adl.org/resources/reports/gab-and-8chan-home-to-terrorist-plots-hiding-in-plain-sight

Buzz Blog Box. (2020, February 2). What is Training Data: Its Types and Why it is Important? Becoming Human: Artificial Intelligence Magazine. https://becominghuman.ai/what-is-training-data-its-types-and-why-it-is-important-f998424c3c9

Calvert, C. (2006). Hate speech and its harms: A communication theory perspective. *Journal of Communication*, *47*(1). https://doi.org/10.1111/j.1460-2466.1997.tb02690.x

Cherry, K. (2020, September 18). How Does Implicit Bias Influence Behavior? *Verywell Mind*. https://www.verywellmind.com/implicit-bias-overview-4178401

Chotiner, I. (2019, July 5). The Underworld of Online Content Moderation. *The New Yorker*. https://www.newyorker.com/news/q-and-a/the-underworld-of-online-content-moderation

Dawson, C. (2020, July 7). Context reduces racial bias in hate speech detection algorithms. *ScienceDaily*. https://www.sciencedaily.com/releases/2020/07/200707113229.htm

Guthrie, E. (2018, October 29). Gab: Its History and Influence in the Tree of Life Shooting. The Pitt News. https://pittnews.com/article/137102/news/gab-its-history-and-influence-in-the-tree-of-life-shooting/

Implicit Bias. (2020, January 3). *Workplace Strategies for Mental Health*. https://www.workplacestrategiesformentalhealth.com/resources/implicit-bias

Koch, Dr. A. (2020, October 14). Online White Supremacy: Looking for a Place to Spread Hate in the Age of Multiple Communication Platforms. *GNET*. https://gnet-research.org/2020/10/14/online-white-supremacy-looking-for-a-place-to-spread-hate-in-the-age-of-multiple-communication-platforms/

Lavin, T. (2018, January 7). The Neo-Nazis of the Daily Stormer Wander the Digital Wilderness. *The New Yorker*. https://www.newyorker.com/tech/annals-of-technology/the-neo-nazis-of-the-daily-stormer-wander-the-digital-wilderness

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLOS ONE, 14*(8). https://doi.org/10.1371/journal.pone.0221152

Ricee, S. (2021, May 26). Subconscious vs Unconscious: The Complete Comparison. Diversity for Social Impact. https://diversity.social/unconscious-vs-subconscious/

Subconscious vs Unconscious Mind. (n.d.). *Diffen.Com*. Retrieved March 11, 2022, from https://www.diffen.com/difference/Subconscious_vs_Unconscious_mind

Vaidya, S., Basu, S., Naderi, A., Wohn, D. Y., & Dasgupta, A. (2021). Conceptualizing Visual Analytic Interventions for Content Moderation. 2021 IEEE Visualization Conference Short Papers (VIS). https://doi.org/10.1109/VIS49827.2021.9623288