The Contribution of Machine Learning Algorithms to Radicalization

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Andre Knocklein

Spring 2020

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

Advisor

Joshua Earle, Department of Engineering and Society

The Contribution of Machine Learning Algorithms to Radicalization

1. Introduction

An event in 2017 served as the culmination of many years of increasingly dogmatic rhetoric. The Unite The Right rally in Charlottesville was a mass gathering of neo-nazi, white supremesist, alt-right, and other extreme right-wing ideological groups that made national news. It was a rally organized with the purpose to protest plans to remove a statue of confederate leader Robert E. Lee. This particular point of contention of the removal of monuments to confederate leaders had become a popular rallying point in right-wing spaces online and many personalities in those spaces parroted talking points that would then be used to organize the Unite the Right rally. This exposed a trend that had been building up for years. Many of the inflammatory ideas that are used by ideologically extreme groups are born in online spaces. This is partly due to the nature of increased connection that the internet provides where geographically distant people can form groups with a large number of members regardless of how extreme the views of those groups are because with enough people statistically there are others that believe the same thing.

What shocked many in the public after the Unite the Right rally was how easy it was for these groups to organize and spread their message and how similar their talking points were to talking points that were being recommended to them in the content they were consuming. Extreme ideological rhetoric can be spread by alluding to the actual point so as to not sound too extreme while people that watch consume that content and are part of these groups know what the content actually means. This content that alludes to more extreme rhetoric is also a way for these ideologies to subtly spread their message to new people who are being influenced without

knowing. It is this type of content that is common even on popular platforms like YouTube and it can lead to what has been coined the Alt Right Pipeline.

The Alt Right pipeline is the phenomenon of online content leading towards more and more radical, right-wing rhetoric. Generally, the way this works is when a user comes across seemingly harmless content that hints towards more radical ideological points. After consuming this content, the user might already be in the pipeline. The channels and content creators that create this type of content "serve as gateways to fringe far-right ideology" (Ribeiro, 2020). The Alt Right pipeline has most closely documented on sites like YouTube which have algorithms designed to serve content to their users.

A core part of this Alt Right pipeline is how these content-serving platforms can recommend content to their users that exposes this rhetoric. In this paper, I will discuss the extent to which machine learning algorithms used by content-serving companies influence radicalization in their users. Every year, society becomes more and more hooked on internet content and social media (Pew Research Center, 2022). This gives the companies that serve that content immense power to influence their users. When combined with the goal of maximizing the time users spend on these websites and the introduction of machine learning algorithms that are so secretive and complex that there is most likely not a single person that completely understands them, the radicalization of some groups of users on these platforms has been observed. To discuss this radicalization is more important now than ever because the political landscape, especially in the United States, is becoming more and more polarized, and discussing a potential cause is important. The most common form of this radicalization has been the aforementioned 'Alt Right Pipeline' and this report will discuss how prevalent it is, what the entry points to it are, and what groups are most likely to fall into it.

2. Research Question

The technology at the center of this STS report is machine learning. Machine learning is a growing field in computer science which is concerned with using data to improve performance on a given task. Specifically, I will examine the usage of machine learning in creating algorithms for content-serving websites. These websites use these algorithms to recommend content that is thought to be relevant to the given user by using the data of what that user has consumed previously and what similar users have consumed in order to recommend new content or tailor a search query to that user. Since these algorithms use data, in many cases without supervision from humans, in order to improve themselves, the actual nature of how they work or what they will produce is hidden. For example, YouTube uses a Deep Neural Network which is a machine learning algorithm which is a complex algorithm which, at the given scale of data that YouTube uses, cannot be supervised (Covington, 2016). The term supervised in machine learning means that there is a human guiding what the algorithm does, in this case, there is no feasible way for a person to guide the algorithm because the amount of data it combes through is much too large. Therefore, this algorithm runs by itself, and people have some high-level understanding of how it works, but the exact reasoning as to what it does is either not known or difficult to figure out. This paper will go into more detail on this in the Technical Analysis Section. Combining this with the fact that companies are very secretive about the algorithms they use makes the internal working of the algorithms impossible to know, but the consequences they create can be readily observed.

My research question is to what extent do these algorithms contribute to the radicalization in political ideology of the users of the websites that use them. I will discuss the

nature of this radicalization as well as how it typically gets started for users. I will also discuss how this radicalization has been coined the “Alt Right Pipeline”. With an ever-increasing portion of the population using the websites that use these algorithms and an ever-increasing sense of polarization (Geiger, 2021) and radical rhetoric, it is important to research one of the ways in which this polarization gets started in those websites. Beyond just the knowledge of how this radicalization works, it is important to realize its harmful effects on the groups of people targeted by radical thought.

3. Platform Analysis

This section of the paper discusses how these algorithms came about at the different platforms and eras of their use. In doing so, I uncover how the landscape of this problem has changed over time and how it might evolve in the future. This is important because there are distinct times when different platforms were popular and hot-beds for the type of radicalization which this paper is interested in.

3.1 Methods and Frameworks

In order to do this platform analysis, some methods and frameworks will be useful. As this paper covers the different platforms and eras, I use case studies as illustrative examples. The entry-points for radicalization have changed over time and the content-creators have also

changed which these case studies can illustrate. With these case studies, I analyze the discourse and rhetoric that is popular with the personalities that are commonly associated with radicalization. I will also analyze this rhetoric in order to understand how the content influences the nature of discourse outside of online spaces of the topics that are prevalent in the recommended content.

The set of case studies that I analyze in this paper are all chosen to represent a specific part of the radicalization pipeline. My analysis of these case studies mainly follows the terminology and analysis framework given by Manoel Horta Ribeiro and co-authors (2020) which outlines how to classify the cases and how to analyze their relationships with each other. Ribeiro uses data collected on YouTube to categorize YouTube channels into categories of Intellectual Dark Web (IDW), Alt-Light, and Alt-Right. IDW channels are those that discuss controversial topics without overtly making extremist points. Alt-light channels talk more openly about and endorse right-wing conspiracy theories. Alt-right channels are those that overtly endorse extremist views and methods. The case studies this paper considers are the Aryan Nation Liberty Net as an example of an early example of extreme rhetoric on the internet, Stormfront as an example of the shift toward user-generated content that radicalizes, PragerU as an example of a YouTube channel that gives users a pathway to more extreme content, and Stephen Crowder as an example of some of that more extreme content. These case studies provide a clear view of both how the pipeline evolved and how users can fall into it.

3.2 Results

3.2.1 Early Internet:

An extremist, right-wing presence has existed on the internet since its inception. In the early internet, when the internet was just “dial-up”, bulletin board systems (BBS) were used by several far-right organizations. One prominent example is the Aryan Nation Liberty Net which dates back as far as 1984 which was used to spread dogma and information about the Aryan Nation (Conway). There were several of these systems set up where geographically distant individuals with similar and extreme ideas could find each other and coordinate. With the limited reach of systems of this type especially compared to the reach modern tools have, they were not generally used as recruiting tools, but could still be used to radicalize their users with propaganda.

Once the World Wide Web was created, the new technology was increasingly adopted by the far right allowing a greater spread of its ideology. An example of this is Stormfront which was a massive hub for radical right wing information on the internet. This was often used as a recruiting tool for other right-wing spaces online. Websites like Stormfront included forums where users could radicalize each other and marked the start of user-generated content (in this case forum posts) leading to radicalization.

A common through-line to right-wing content and ideology is the creation of enemies and scapegoats. This will be evident when looking at more platforms and content-creators, but can be seen with the website in the early internet. Both of the above examples, have tight connections with white-nationalist ideas which clearly scapegoat people of color for problems they might see in their own lives or society. These early websites are very explicit with their intentions and

ideology, but more modern content-creators are more subtle but still try to lead their users to the same conclusions.

These older eras of the internet serve as a good foundation for the discussion of how the new technology of content recommendation algorithms might inadvertently lead to the result which the creators of those early websites would have hoped their websites could. Moving towards discussion of social media companies, it is important to understand the context in which extremist content occupied the early internet.

3.2.2 Modern Internet:

While this paper mostly focuses on trends on YouTube, it is important to understand that this problem applies to other platforms too. These other websites also use algorithms to serve content to their users, so they are useful to consider. Twitter has been linked to promoting right-wing extremist content and it seems to be growing. There has been an increase in the following of white-nationalist accounts of "more than 600% since 2012" (Berger).

This time frame (2012 to 2016) is also when there was an explosion of right wing viewership on YouTube. The increased political polarization between the 2012 election and up to the 2016 election spurred the creation of highly political content on YouTube. This timeframe saw the election loss of the Republican party in 2012 and a reactionary galvanization afterwards. In the rest of the four years, the rise of populist leaders like Donald Trump ahead of his election win in 2016 increasingly pushed the overton window on what is acceptable in the campaign process to greater extremes. This change in the overton window subsequently fueled the

extremization of online rhetoric. This can be seen by the drastic increase in the level of interaction (likes and comments) by viewers on videos that could lead to radicalization (Ribeiro, 2020). This is also the time frame when this issue became known to most people and when most of the research into this topic began.

When discussing radicalization pathways on YouTube, it is important to distinguish between types of content. Blatantly radical content is termed as "Alt-Right" content, and content that might lead towards radicalization is termed as "Alt-lite" and "Intellectual Dark Web (I.D.W.)" content. What distinguishes these groups is that IDW content "discuss[es] controversial subjects like race and I.Q." while alt-light content flirts more with right-wing conspiracy theories like the great replacement. Alt-right content explicitly "sponsor[s] fringe ideas like that of a white ethnostate." (Ribeiro, 2020). Research has found that the overlap between active users in these spaces on YouTube is substantial which gives evidence towards the fact that alt-lite and IDW act as gateways to alt-right content.

The Ribeiro study highlights channels that are examples of each of these groups, and these are good case studies to consider. In the IDW category is an influential channel called PragerU which stands for Prager University which is not a university. It is a conservative propaganda channel created by radio talk show host Dennis Prager.

PragerU is a perfect example of a channel that masks its extreme ideas in order to appeal to as many people as possible in order to radicalize some of them. It pretends it is an academic institution that discusses modern issues from an independent viewpoint. This is not the case. PragerU's content takes the form of different right-leaning presenters making short video presentations about a topic. This means that PragerU is a tool for these presenters to build an audience on content that many might not immediately suspect has an agenda to radicalize its

users because it “lends this veneer of legitimacy to its presenters and the ideas they express” and “portrays those presenters as unbiased moderates too.” (Dickinson, 2021). PragerU acts as a hub for the IDW and an interface between mainstream politics and the IDW and alt-lite. PragerU has had mainstream political pundits like Tucker Carlson and Candace Owens as presenters which increases its credibility to the public, but also its scope and capability of radicalization. In addition to these mainstream pundits, it has also had many presenters who are part of the alt-light and IDW groups like Steven Crowder, Charlie Kirk, and Ben Shapiro.

Steven Crowder, for example, is a Canadian-raised political commentator whose main source of fame was his debates with college students where he challenged them to change his mind. This tactic of being prepared for a debate to have against young adults who were not prepared to have a debate worked to gain him many followers and pushed the narrative that American universities are breeding grounds for left-wing ideas and are inhabited by young people who don’t know how to support their ideas. He has his own show which has played host to conservative figures like Donald Trump and Ted Cruz. The topics discussed range from denying human-caused climate change to bashing trans people.

Most of the presenters on PragerU and other alt-lite and IDW channels don’t explicitly platform content from the alt-right. They tend to use alt-right dog whistles and allude to alt-right talking points without explicitly stating them in order to appeal to alt-right viewers. The other effect of the allusions to alt-right talking points is that some viewers might extrapolate those allusions and arrive at the talking point which leads them to the alt-right. Take the example of the great replacement theory. This theory states that minorities and people of color and immigrating to majority white countries in order to replace the white people in those countries (Global Project Against Hate and Extremism, 2022). IDW and alt-light content will allude to this theory by using

dog whistles like that immigrants are destroying the culture of those countries. This plants the seed in users and can lead to more overtly prejudiced content and rhetoric like that used at the Unite the Right rally where “Jews will not replace us” was chanted.

To summarize, the radicalization pipeline on YouTube is organized as follows. IDW channels are mainly contrarians that appear to push against the system and portray themselves as neutral. They will often platform alt-lite channels or personalities who use talking points similar to those used by the alt-right which then exposes users to those talking points until those users are comfortable with those points and will consume alt-right content without questioning it.

3.2.3 TikTok:

The new social media platform that seems to be afflicted with the same pipeline problem is TikTok. TikTok also uses an algorithm to create its “For You” page, but because the nature of the content on TIkTok is incredibly short-form, this algorithm must recommend videos incredibly rapidly. This causes an even faster feedback loop. The main issue that is leading towards right-wing extremism on TikTok is that of anti-trans content. Being anti-trans is still widely accepted in mainstream society, so this anti-trans content could be equated to IDW or alt-lite content. When interacting with anti-trans content, TikTok recommends many more overtly extremist videos including calls to violence and white supremacy (Media Matters, 2021). The same pattern of increasingly extreme content can be seen as the more anti-trans content is watched, the more extreme other videos become.

4. Technical Analysis

4.1 Deep dive into the mechanics of the algorithms

In order to truly understand the causes behind this problem, I think it is important to understand, at a basic level, the mechanisms behind how these algorithms work. In order to do this, I will use YouTube's algorithms as an example. Most of this information comes from "Deep Neural Networks for YouTube Recommendations" by Covington, Adams, and Sargin. They work on the recommendation algorithm at YouTube and have laid out a good overview of how it works.

As mentioned before, YouTube uses a Deep Neural Network (DNN) machine learning algorithm to serve recommendations. A Neural Network is simply a model which has layers of nodes which are connected to nodes on the next layer. When an input is passed in, it runs through the model and creates an output. This output is then compared to a goal which then causes certain connections to update their values which affects the math that is done on later data. Basically, it is just a model for creating many mathematical equations. The way these models become useful is by running a lot of data through them which then updates the values of the connections so that the model can learn to recognize patterns. A DNN is any neural network that has many layers of nodes.

YouTube's algorithm actually uses two DNNs to recommend videos. Both of these algorithms use watch time as the goal. That means they are going to try to recommend videos that maximize watch time by using the videos that a user spent a long time watching as

examples. This means that the " explicit feedback mechanisms" that exist on YouTube like dislikes are not used to train the model, but watches and watch time is more important.

The first DNN takes in millions of videos and outputs hundreds of videos that might interest the user. This is termed "candidate generation". In order to do this, it uses the user's "history and context" which concerns what the user had previously consumed or searched for as well as what other, similar users have consumed or searched for. This similarity is " expressed in terms of coarse features such as IDs of video watches, search query tokens".

Only in the second DNN, which is termed the ranking model, are any of the features about the video itself taken into account. It also takes "history and context" as input like the previous step. This DNN takes the hundreds of candidates from the "candidate generation" DNN and ranks them by how likely the user is to watch them. This is ultimately the algorithm that creates the recommendation page that users see.

This is only the algorithm used by one company to serve content, but since YouTube is the largest of these companies, it is a good example to use. Other content-serving platforms most likely make their algorithms based on YouTubes example. Of course, they will be slightly different and their goals might be slightly different but the idea is the same. The point of these algorithms is to keep users on the platforms as long as possible and keep them coming back by serving content that the users will be likely to watch.

4.2 Results of the deep dive

This technical look at the algorithms provides some insights into why an extremization pipeline like the alt-right pipeline can manifest. One aspect of the design that can lead to the extremism pipeline is that the algorithm is partly based on what similar users might be likely to watch. Another aspect of the design that could be leading to this problem is the goal on which the model is trained is purely based on watch time. These two factors lead to the radicalization pathway purely based on the architecture of these algorithms and not their implementation.

The first reason mentioned above is a problem because once one user in a demographic is radicalized, others in that group are more likely to also see the content that is recommended to the radicalized user. This means that if a given user is interested in a topic that a radicalized user is also interested in, the algorithm might think both could be interested in radical content. For example, let's assume that a user is interested in woodworking and watches many woodworking videos. Let's also assume the demographics of users that watch woodworking videos tends to skew right and many of them watch right wing political pundits. YouTubes algorithm might now recommend a right wing political pundit to the user just because they watch woodworking videos. This example is not out of the realm of possibility, and I am certain that similar situations have happened.

The other problem is that the goal is purely based on watch time. This is mostly a shortcoming of machine learning models in general. They need a clear, mathematical way to determine how accurate their predictions are. There is no way that they could optimize for nebulous outcomes like learning or betterment of the users. That being said, the fact that they ignore user input in their calculations can lead to problems. Even if a user watched an entire video in order to understand it before disliking it due to disagreement for example, that video will still count as a positive example as to what to recommend to that user next. This leads to

inflammatory content that is easy to watch even if users disagree with it to be pushed to many people. This type of content seems to be one of the major entry points for the radicalization pathway.

This analysis is not perfect, and there are many more details about the algorithm that YouTube uses and there are many companies whose algorithm I haven't touched on, but this trend seems common in many of the platforms that have this problem. These algorithms are difficult to change because they have been running and collecting information for so long which seriously discourages their parent companies from making any substantial enough changes to solve these problems.

5. Combined Analysis

With the knowledge of both how the right-wing space on the internet was created and is organized and how the algorithm on a platform like YouTube works, it is possible to create a full picture of how this radicalization works.

5.1 How the goal of the algorithm shapes content

As was already mentioned, the goal of the algorithms used by these platforms is to maximize the time that users spend on the site. The best way to do this is to have a good method for recommending content that users are likely to watch, regardless of what that content actually

is. This creates the situation where possibly problematic videos can be watched which is feedback to the algorithm to recommend more of that type of content which creates a feedback loop towards more and more problematic content. This is how consuming IDW content could lead towards alt-light content which then leads to alt-right content without the users explicitly seeking out that content.

The overlap between the IDW, alt-light, and alt-right also creates the environment necessary for the algorithms to recommend content of one of those categories when a user consumes content from another. The fact that the algorithms take similarities between users into account when recommending videos means that this overlap is a perfect tool for influencing users to become more extreme. Users that are already in the alt-right sphere tend to consume alt-light and IDW content, so when the algorithms consider a user that consumes only IDW or alt-light content they can assume that user will also be interested in the more extreme, alt-right contend. This is the primary mechanism of the pipeline.

5.2 Current state of the problem

The radicalization pipeline on the internet due to algorithms is still alive. A recent example of its effects can be seen in the popularity of misogynistic personalities like Andrew Tate gaining popularity through their online presence. Tate is a controversial personality and influencer who is known for his views on relationships which tend towards misogyny. He is also known for being at the center of a sex trafficking investigation in Romania where he immigrated to due to facing sexual assault investigations in other countries. He is part of a new movement of

lifestyle and finance influencers that spout outdated views on relationships and women. Much of this particular example had to do with fan accounts uploading clips to TikTok and YouTube which became viral because they drove engagement and appealed to individuals in partly-radicalized spaces online.

Most of the research used in this paper has also been recent which is evidence that this problem is only now being fully understood and that it is still one that needs focus because it is still prevalent. The trends mentioned, even those from the early internet are still present. Websites like Stormfront still exist and so do forums and discussion boards that can be linked to by content served by algorithms which also speeds up radicalization. The algorithm will keep recommending the way it has been because it is effective, and it is easy to fall for the propaganda and priming that this type of content provides.

6. Discussion

6.1 Responsibility Analysis

The blame for the trend of radicalization should not be placed squarely on the shoulders of those who fall into the pipeline. In many ways, they were manipulated by forces that they could not control and that were hidden from them. The algorithm just does the math that it has found works the best to keep users on the site, and since the technology is advanced, that math tends to be accurate to accomplish its goal and users therefore tend to watch the videos it recommends. There are undoubtedly instances where every user that fell into the pipeline could climb out of it but chose not to, but making that change requires willpower that is hard to derive

especially when that user has been shaped by the content they watched over a long period of time.

More of the blame should be placed on the content-creators that feed off of the algorithm and profit from users falling into the pipeline. I would argue that most of the content-creators in the alt-lite and IDW spaces are fully aware of the views of their consumers and the direction their consumer's ideology gets pushed over time. In many cases, the content-creators seem to actively encourage further exploration into extreme ideologies. This is most likely due to the fact that the more extreme users will still agree with most of their points but are much less likely to be swayed to the other side of the political spectrum thereby removing a user. It is also important to keep in mind, however, that content-creators tend to adapt their content so that it fits in line with what the algorithm recommends because recommendations are an important source of viewers for these channels. It can seem like a chicken-and-egg kind of problem. Is the algorithm shaped the way it is because of the content that users tend to watch, or is the content users tend to watch the way it is because of what the algorithm tends to promote?

Ultimately, the only party in this equation that has any real power to stop this trend is the platform itself. Even though the algorithm itself is only loosely in control of the platform itself, the platform can create moderation systems to vet the steps in the pipeline that connect users to more and more extreme ideologies. The goal of the platform is purely monetary, so an intervention like this is unlikely, so for the time being the best thing to do is put pressure on the platforms where this is common and spread the information of this phenomenon so that fewer people fall into it.

6.2 Conclusion

This pipeline which has been made worse by content-serving algorithms is one of the primary reasons why extremism seems to be on the rise. The ease of access to extremist sources of propaganda on social media platforms is alarming. More research and attention has been brought to this problem and there have been efforts by other content-creators to attempt to combat this phenomenon. It is unlikely that these efforts will be successful because of how polarized the discussion of politics has become and someone who might lean right might be unlikely to even consider the points made by someone trying to counter the pipeline. It is still good to push back because even the few users for whom it might be possible to reverse the effects of the pipeline are important to spare from its effects.

Berger, J. (2016). Nazis vs. ISIS on Twitter: a comparative study of white nationalist and ISIS online social media networks. *GW Program on Extremism*.

Bilewicz, M., & Soral, W. (2020). Hate Speech Epidemic. The Dynamic Effects of Derogatory Language on Intergroup Relations and Political Radicalization. *Political Psychology*, *41*(S1), 3–33. https://doi.org/10.1111/pops.12670

Covington, P., Adams, J., & Sargin, E. (2016). Deep Neural Networks for YouTube Recommendations. *Proceedings of the 10th ACM Conference on Recommender Systems*. https://doi.org/10.1145/2959100.2959190

Dickinson, R., & Cowin, T. (2021). The Kids Are Alt-Right: An Introduction to PragerU and Its Role in Radicalization in the United States. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4187075

Dunbar, E. W. (2022). *Indoctrination to Hate: Recruitment Techniques of Hate Groups and How to Stop Them*. Praeger.

Geiger, A. (2021, April 9). *Political Polarization in the American Public*. Pew Research Center - U.S. Politics & Policy. https://www.pewresearch.org/politics/2014/06/12/political-polarization-in-the-american-public/

Global Project Against Hate and Extremism. (2022, September 30). *The Great Replacement*. https://globalextremism.org/the-great-replacement/

Lewis, B. (2018, September 18). *Alternative Influence*. Data & Society. https://datasociety.net/library/alternative-influence/

Lyons, M. N. (2017, January 20). *Ctrl-Alt-Delete: The origins and ideology of the Alternative Right*. Political Research Associates. https://politicalresearch.org/2017/01/20/ctrl-alt-delete-report-on-the-alternative-right

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism* (Illustrated). NYU Press.

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. a. F., & Meira, W. (2020). Auditing radicalization pathways on YouTube. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3351095.3372879

Social Media Fact Sheet. (2022, November 16). Pew Research Center: Internet, Science & Tech. https://www.pewresearch.org/internet/fact-sheet/social-media/?menuItem=3814afe3-3f3c-4623-910b-8a6a37885ab8

TikTok's algorithm leads users from transphobic videos to far-right rabbit holes. (2021, October 5). Media Matters for America. https://www.mediamatters.org/tiktok/tiktoks-algorithm-leads-users-transphobic-videos-far-right-rabbit-holes