ABSTRACT

Title of Thesis: LEMMA: A Data-Driven Approach to Modeling the Spread of Extremism Over Online Platforms

Thesis Directed by: Professor Pierre-Emmanuel Jabin
Department of Mathematics and Huck Institutes
Pennsylvania State University

The online spread of extremist ideas has been a growing problem. Team LEMMA has worked to quantitatively model the spread of extremist ideas over Reddit in order to gain insight into how they may spread. A modest dataset of Reddit comments were manually rated on the level of extremist rhetoric present and used to train a machine learning algorithm to automatically classify large swaths of Reddit data. These ratings were then fit to a predictive agent-based model with the hopes of better understanding past trends and potentially forecasting future spread of extremism.

LEMMA: A Data-Driven Approach to Modeling the Spread of Extremism Over Online Platforms

by

Mitchell Fream, Nathan Hayes, Sahil Kochar, Paul Kolbeck, Charlie Schneider, Russell Schwartz, Olivia Sharon, Yuang Shen, Winslow Weiss, Robert Wolle

Thesis submitted to the faculty of the Honors College of the University of Maryland in fulfillment of the requirements for a Citation from the Gemstone Program, 2022

Acknowledgments

We would like to thank Dr. Pierre-Emmanuel Jabin for his mentorship and support throughout the past three years. We would also like to thank the Gemstone faculty and staff, who have been with us throughout our tenure in the program. Finally, we would like to thank our discussants, Dr. Derek Paley, Dr. Philip Resnik, Dr. Alan Tsang, Dr. Steve S. Sin, and Rhyner Washburn.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Social media has become a central method of sharing ideas in society. However, extreme ideas also spread over social media and can produce harmful outcomes. Radicalization on the internet has produced real world consequences, including lone wolf terrorist attacks. Social media websites attempt to use content moderation to limit the spread of extreme ideas on their platforms, but current methods of content moderation require manual or imperfect automated systems that must scale with the ever-rising number of users. In order to address this problem, our project attempts to measure and mathematically model the spread of extremist ideology over Reddit.

1.1 The Danger of Extremist Ideas on Social Media

There are several key features of the spread of extremism over social media that are relevant to our project. The first is that radicalization can occur through direct contact with extremists, through a group setting, or through self-radicalization. We will refer to the first kind as deliberate radicalization. There is clear evidence of terrorists actively and directly attempting to radicalize people using propaganda spread over social media [1]. Next is radicalization occurring in a group. Being in a group can cause radicalization if the group as a whole moves towards increasingly extreme ideas as like-minded people gather and spread information [2]. Finally, there is self-radicalization, which is when people are radicalized by extreme ideas on the internet but not

through direct contact with radical individuals.

Self-radicalization can lead to "lone-wolf terrorism," where individuals perform violent action after being radicalized in this manner. Examples of this phenomenon include the perpetrators of the Boston Marathon massacre, the Pulse nightclub shooting, and many more [3]. There is debate on the ways the internet facilitates self-radicalization. Some researchers suggest that the internet normalizes extreme attitudes unacceptable to share offline, creates echo chambers, and fosters group polarization. Others suggest that the internet is not a direct cause, claiming that self-radicalization is actually more due to social alienation and other individual factors [3].

Another factor that can contribute to being radicalized is age. There is evidence that youths are specifically targeted by radical organizations. Extremists directly recruit youths with materials targeted towards them, such as cartoons, video games, and chat rooms [4]. The U.S. Department of Homeland Security suggests that there are three primary ways youths are radicalized. One: they encounter radical content when searching for entertainment, two: they encounter radical content when seeking information on culture and traditions associated with a radical group, and three: they find an extremist group when trying to find a group to identify with [4]. These key features on how extremist ideas are spread, who spreads them, and who is vulnerable to them will inform our model.

A detailed model describing the spread of extreme ideas could help us to understand how agendas spread, what sort of targeting strategies are effective, why subgroups form, and even help us to predict future trends.

1.2 Introduction to Reddit

Social media platforms provide an environment for users to share and trade ideas online, including extremist ideas. Reddit is a particularly useful platform for studying ideas because its structure clusters ideological groups together and discussions are almost always in text format. In addition, Pushshift.io provides an API for accessing all posts and comments made on Reddit [5]. Our team made use of its archives to download the relevant months of posts and comments to a large storage drive.

The environment that ideas exist within will influence how users discover and discuss those ideas. Reddit's structure is based on subreddits, communities where posts generally fall under a single topic. Within each subreddit, users can create posts which are composed of a title, and a body filled with text, a link to an external site like a news article, or visual content like an image or video. Users may then comment under the post with additional text content, and users may add comments to other comments, creating branching conversations under the original post. More information about Reddit may be found at the company's about page [6].

1.3 Social Media Moderation

Social media platforms usually moderate content for a variety of reasons including cultivating a particular kind of environment, appealing to advertisers, and reducing the spread of unethical, harmful, or potentially illegal content on their website [7]. Many websites used to use community-level methods for content moderation where users received punishments from other users in the same community. Some users may be given special moderation privileges that allow them to

deliver punishments such as suspensions to users who violate content policies.

Reddit employs this method; moderators are usually members of the community they moderate. Subreddits enact specific rules and policies about what content is allowed in each community, but Reddit also has site-wide rules. Users can flag or report posts and comments in accordance with either subreddit or site rules, but moderators ultimately make the decisions for what content is allowed in the community. The highest moderation privilege level is still held by site administrators.

Popular social media sites have grown too large to be effectively moderated by a select group of people, and many modern platforms are now structured around curated feeds as opposed to more isolated communities. Moderators can no longer only monitor their own community; moderators on sites like Twitter operate site-wide. Due to the COVID-19 pandemic, Twitter migrated to almost entirely automated moderation after laying off human moderators [8]. Automated approaches of large-scale social media moderation are common. Most platforms use a pattern-matching algorithm to identify content that is similar to previously flagged violations, which is not a form of machine learning [8].

In addition to the feasibility concerns of human moderators, there is also a question of ethics. Human moderators are subjected to content that might violate the policy of the platform, which can include psychologically troubling material.

In producing a measurement method and predictive model of extremist ideas, we hope to provide effective automated tools for the moderation of extremist content on social media platforms and identification methods so that platforms may be more aware of how extremist ideas behave in their environment.

1.4 Previous Approaches

Extremist ideas that spread over social media can cause destructive events in the real world, and their spread can be tracked by the presence of extremist language in social media posts. The most simple method of identifying extremism on social media is to create a list of extremism-related keywords and flag any content containing a word on the list [9]. This rudimentary approach has some issues, namely the fact that this blocklist must be manually curated and the exact matching can be easily sidestepped with simple substitutions (e.g. hiding "kill" as "k!ll") or misspellings. Additionally, these manually-curated blocklists reflect the current knowledge of the moderating party, and therefore may not accurately represent language used by extremists. More sophisticated methods use mechanisms like regular expressions to catch substitutions and simple word structures, but they generally under-perform in identifying extremist content.

Machine learning has offered promising improvements to extremism identification. Broadly speaking, machine learning methods fall into two categories: supervised and unsupervised [10]. Supervised learning involves training a machine learning classifier on a dataset along with labels that identify each element of the dataset as extremist content or not. These labels are often manually generated, and the classifier attempts to learn patterns from the labeled dataset and generalize them to classify unlabeled data. Unsupervised learning involves training a classifier on a dataset without any labels, allowing the algorithm to separate the data into a specified number of classes according to some heuristic. The classes generated by this kind of learning are not necessarily quantifiable by succinct labels. However, machine learning models are infamously susceptible to bias, especially in the dataset. Both the construction of the dataset and the labeling of the dataset can introduce human bias that the machine learning algorithm may incorporate into

its classification [11]. Examples of possible bias in dataset construction include the geographical distribution of where content in the dataset originated and the balance between extremist and non-extremist content. Manual labeling is another source of possible bias, as generally human operators must decide the labels of each element, possibly introducing their own bias into the algorithm.

Natural Language Processing (NLP) is another recent approach to extremism identification that attempts to use machine learning algorithms in conjunction with analysis of semantic and syntactic elements of textual data [11]. NLP algorithms generally extract these textual features to be used as input vectors for a machine learning classifier. By identifying text-specific features in the pre-processing of the data, NLP aims to guide the classifier to make more accurate predictions. However, since NLP relies on machine learning, it is also susceptible to the same sources of bias, so the dataset and labels must be constructed with considerable care towards reducing possible bias.

Sentiment analysis uses NLP to determine a text's author disposition to the content of the text. Sentiment analysis has been used to help identify whether a poster on social media belongs to the incel community, but this method requires additional textual factors to determine whether a negative sentiment indicates extremism [12] such as the keywords described previously.

The previously discussed methods all attempt to identify extremist content, but none attempt to track the spread of such ideas. Previous works in tracking the spread of information take a number of approaches. One such approach is topic tracking, which models ideas as "topics," generally represented as a collection of associated words with a probability distribution that approximates how common they are in the corpus [13]. Sophisticated topic tracking models such as TIDE [14] incorporate a model of information diffusion through a social network, approximating

the dynamics of social media. The COEVOLVE [15] model is not a topic tracking model, but it simulates how information diffusion through a network can then change the network and vice versa, providing an approximation of social network dynamics.

1.5 Defining Extremism

In order to measure and model the spread of extremism, we must first define it. First, it is key to distinguish between extremism and radicalism: radicalism is an ideology characterized by political ideas outside the norm of society, while extremism additionally views violence as "a legitimate form of political action" [16]. This distinction is key when we look at extremist ideas, which indicates ideas that reflect and promote an extremist ideology. A further note is that we are looking at extremist ideology as a whole rather than features of particular extremist groups. Rather than attempting to track extremists or ideology related to particular extremist groups, we decided to measure extremism through features of extreme ideas in general. This was largely due to the potential bias and ethical issues that might arise from us attempting to measure or track specific extremists or groups with automation. However, the literature on extremism tends to focus more on these measurable features of specific groups, so we supplemented what we could find with literature about extremism that spoke to extremist ideology in general.

1.6 Our Approach

As stated above, the goal of our project was to mathematically model extremist ideas on Reddit. In order to do so, we broke this down into two problems: measuring extremist ideas in a large dataset, and building a mathematical model using this data.

This first task of measuring extremist ideas was broken down into two subtasks: developing a manual measurement for extremist ideas and automating that measurement. These two tasks were developed hand-in-hand. After an initial, categorical rating guide was developed with the intention of using logistic regression for automation, we tested and refined it over rounds of rating comments. During this process, we discovered that our initial approaches of categorical rating and logistic regression were unsuitable for the complexity of the rating task. This led to the development of a revised rating guide, which redistributed the initial categories across a scale allowing for more flexibility, in conjunction with the switch to using machine learning, specifically the BERT Language Model, for automation.

For the second task, once we obtained our reliable method of measuring extremism, we sought to use this data to construct a predictive model of the behavior of extremist Reddit communities. We turned to agent-based models of consensus formation, and modified one of these to accommodate our particular context [17]. We trained this model on data drawn from our BERT automated rating using gradient descent and regressive techniques. Using this model, we built a more sophisticated population-level model which aggregated our agent-based results and examined the subreddit as a whole. Finally, we performed tests to evaluate its predictive capabilities.

Our finalized method of achieving our goal is outlined in Table 1.1, while future chapters detail the process by which this method was found and its results.

Table 1.1: Final Approach

Step	Method Used
Data Collection	Downloaded from the Pushshift.io Reddit archive
Manual Extremism Measurement	Rated 2500 Reddit comments according to our revised rating guide
Automated Extremism Measurement	Treated the manually rated comments as training data for the BERT Language Model
Mathematical Modeling	Implemented modified consensus model with aggregate population-level approach to predict change in subreddit extremism

Chapter 2: Data Gathering

In order to properly study the spread of extremist ideas on Reddit, we needed a large, relatively contained dataset that featured extremist ideology. We chose `r/Incels` because it fit these requirements.

2.1 The `r/Incels` Subreddit

`r/Incels` was a subreddit dedicated to the discussion of involuntary celibate ideas which was self described as "for people who lack romantic relationships and sex, but mostly geared towards those lacking a girlfriend or seeking marriage." The subreddit was created in 2013 but had little traction until mid-2016 when the subreddit gained hundreds of subscribers over a period of weeks following the quarantining and subsequent banning of the `r/Truecels` subreddit, a

community with similar discussion interests, for breaking website rules. r/Incels was banned in November of 2017 after gaining over 42,000 subscribers. Much of the most popular content in the community presented severely misogynistic and hateful views of women. Many posts would receive over one hundred comments.

r/Incels was considered an appropriate subreddit to study due to its size and content. Comments under popular posts can evolve into conversations connecting several users but are still likely to be seen by many of the users looking at the post. Posts with much more engagement tend to have hundreds or thousands of comments which will never be seen by nearly all users who view the post. Sampling the comments for a community of this size provides us a useful amount of both neutral and highly ideological comments. The content of r/Incels is generally relevant to the discussion of involuntary celibate ideas and does not contain an unusually high amount of satirical or otherwise deceitful comments. This allows for reasonable analysis of individual comments without excessive context.

2.2 Method

Obtaining the r/Incels dataset itself created some logistical concerns. While fetching comments directly from Reddit was possible, there were a few problems with this approach. Reddit comments which have been deleted are inaccessible through the Reddit API, and because r/Incels was banned from Reddit several years ago, there are many deleted comments. In particular, we wanted to avoid bias caused by more extreme comments being removed by Reddit moderators. Additionally, Reddit limits the bandwidth at which comments can be downloaded, which would have made scraping the entire dataset take a prohibitively long time. For this reason,

we chose to use data from Pushshift, which hosts a freely accessible backup of Reddit data at `https://files.pushshift.io/reddit/`.

Pushshift provides an online search API, which can return comments based on filter criteria including subreddit name, but this API call does not allow for more than a few hundred comments to be retrieved at once. Separately, Pushshift offers compressed files containing all the comments posted to Reddit in a single month. We chose to download these month long archive files for every month until `r/Incels` was banned and create a local database that we could search at will. In order to do this, a new external USB hard drive was purchased to hold an SQL database. We created a python script to automate downloading, unzipping, and parsing the archives from Pushshift and inserting the comments into our local database. After all the comments were downloaded, we filtered for only comments in `r/Incels` and exported them to a CSV file for further use. This resulted in approximately 3,000,000 comments, about half of which could be immediately discarded as obviously bot-generated, auto-moderator-generated, or otherwise corrupted. The comments from other subreddits were retained but not used.

Chapter 3: Extremism: Background and Measurement

In order to model how extremist ideology spreads, we need to develop an approach to measure extremist ideology. For this, we require a case study to develop a tool to measure content for extremism. The `r/Incels` community is suitable for this task because incel ideology is easily identified by key insider language, allowing us to easily discern between innocuous

and extremist content. Investigating `r/Incels` allows us to document the spread of easily identifiable extremist traits in an online community dedicated to incels.

3.1 Background on Incel Ideology

The term "incel" was first coined in 1997 by a Canadian university student hoping to create a community for anyone experiencing "involuntary celibacy," regardless of identity [18]. However, as the community attracted misogynistic men, the original organizer left, and incel ideology was guided by the remaining members. Throughout the 2000s and 2010s, incel ideology became hateful and even violent, with mass murders perpetrated by self-identified incels George Sodini in 2009 and Elliot Roger in 2014 [18]. Incel ideology is best understood as a subset of male supremacist ideology, and community interaction is almost entirely done over the internet in specialized incel communities and websites.

3.2 Justification of Automation

Due to the overwhelming number of comments posted on `r/Incels` during its time online, any attempt to detect and track extremist ideas must either be limited to a small subsection of the community or be automated. With a scope limited to a small subsection of the community, it is possible to fully describe the behavior of extremism in that subsection. Manually determining the level of extremism in hundreds of comments across several posts is tractable.

For the goals of our model, this limited scope is insufficient. Our agent-based model must know how individuals' engagement with extremist content changes over time. Many users post too infrequently for a manual review of even one week of posts to show how engagement changes

over time. Manual review becomes increasingly intractable on posts with higher engagement which could exclude more popular or controversial ideas or posts that have extensive interactions with other subreddits. Any model based on a subset of the community would exhibit strong biases towards the behavior of the subset and could be ungeneralizable to larger communities with different types of interactions.

If we intend our model to describe the behavior of general patterns of extremism across Reddit and include the dynamics of multiple communities interacting, limiting our scope to fully describing a subset of the conversation would not accomplish these goals. This motivates the use of an automated approach on a larger dataset. Automated methods of detecting extremist ideas fall into two broad categories: lexicons, or artificial intelligence. Either method requires a method for identifying extremist ideas.

3.3 Features of Extremists and Extremism

Extremist narratives developed by extremist groups share characteristics that we used to develop our measures of extremism. Many of these are outlined in a paper by Furlow & Goodall [19]. One key extremist narrative is the idea that there is a problem with the world and that violence should be used to solve this problem. Another is that the world is divided into an "us" and a "them"—an ingroup and an outgroup [19]. Extremists are also more intolerant of other groups [20]. These factors tie into an overall narrative designed to inspire violence: blaming a problem with the world on an outgroup and reinforcing the divide between the ingroup and the outgroup, with intolerance feeding into hatred of the outgroup and the creation of a culture of fear, culminating with the justification of violence against the outgroup in order to solve the

"problem" [19]. We characterized more adherence to and acceptance of features of extremist ideology as more indicative of a higher extremism rating.

3.4 Justification for using General Extremism Measure

Research showcases that extreme ideologies can be linked to distinct psychological features [20]. Rather than drawing from any one specific extremist group, we wanted to draw from features and narratives common to extremism as a whole. It was important to us to not develop a system of measurement that exhibited bias against groups, religions, or peoples that are stereotyped as extremist.

3.5 Categorical Rating Approach

The team rated a number of comments from `r/Incels` on whether the following features were present: *character assassination and name calling*, *us vs. them rhetoric*, *war rhetoric*, *intolerance of other groups*, *rhetoric of fear*, *problem with the world statements*, *problem should be fixed statements*, and *calls for violence*. These featural categories were drawn from the narratives outlined earlier.

At first, the categories were judged independently. Each category had an associated numerical extremism value, and the presence of a category feature in a comment added the category's value to the comment's overall extremism measure. Categories which we considered more extreme received a higher associated extremism value. In order to facilitate a simple transition to automated labeling, we attempted to prioritize specific language in defining a category. For example, a comment which used the word "they" to denote an outside group or "us" to denote

members of the subreddit would be assigned the *us vs. them rhetoric* category.

3.6 Modifications to Categorical Rating Approach

After our initial rounds of ratings, revised our approach based on our experiences. In these rounds of ratings, the categories of *character assassination*, *us vs. them rhetoric*, *intolerance of other groups*, and *problem with the world statements* were frequently present. Each of these categories were present in between 8 and 14 percent of the rated comments. However, the other categories—*war rhetoric*, *rhetoric of fear*, *statements that the problem should be fixed*, and *calls for violence*—were much more rare, each showing up in less than 2% of the comments. The infrequency of these categories would make them difficult to properly analyze, so we revised them for greater applicability. The *war rhetoric category* in particular was only found in comments that incorporated other categories, so the category was removed altogether. In addition to its rarity, the *rhetoric of fear* category was too vague to be identified consistently. With this in mind, we merged it into *intolerance of other groups* as a new category, *hate speech*. We considered the last rare category, *calls for violence*, clear and integral to the extremist narrative and kept it regardless of its rarity.

Beyond the extremist narratives themselves, we also found that we needed to measure language resulting from how people reacted to and interacted with these extremist narratives. One category we added was *insider language*. The comments demonstrated an extensive lexicon of insider language, and this terminology was commonly present with other features of extremism. This insider language showcased exposure to and adherence to extremist narratives, and as group membership and identity are an important component of radicalization, we thought it was important

to capture. We also added *profanity* and *harassment* as measures of anger and negative sentiment. Anger and fear towards an outgroup result from hatred cultivated by adherence to the extremist narrative.

3.7 Problems with Categorical Rating Approach

After further rounds of rating comments, we saw that the modified rating system broadly reflected extremist narratives present in the incel community. However, there were numerous problems with the approach.

For one, we faced problems with translating a set of features to a measure of extremism. It was difficult to set the threshold where a comment should be counted as extreme and it was difficult to assign values to the categories that appropriately raised comments above this threshold. We considered using a second layer of values to describe the interactions between categories, but the number of additional values required by such a system would be too difficult to reliably determine manually. It was possible to optimize the values to match our judgements of a comment's extremism, but we felt that such a system would end up overfitting to our limited manually labeled set.

In addition, there were a number of issues with identifying category features within comments. We calculated the Fleiss' κ score for each category (see Table 3.1) to serve as a quantitative measurement of inter-rater reliability, and our best score was in the *calls for violence* category with a value of 0.572. This reflected only moderate agreement according to the standards developed by Landis and Koch [21]. For a less clear-cut category such as *us vs. them rhetoric*, the κ value was as low as 0.275. In order to use our manually labeled data or extend it to automated methods,

we needed more confidence that we could accurately and reliably make judgements on each and every category.

Table 3.1: Category Inter-rater Reliability

Lexicon Category	Fleiss' κ
character assassination and name calling	0.504
us vs. them rhetoric	0.275
intolerance of other groups	0.512
problem with the world statements	0.374
problem should be fixed statements	0.513
calls for violence	0.572
insider language	0.552

One of our primary motivations for using the categorical approach was to map the categories on to a relatively simple lexicon-based approach for automatic rating. When this approach was implemented, we found that it was not accurate enough at matching our manual judgements (see Figure 4.1). Between this and the many problems we had with the manual side of the rating system, it was clear that we would need a more sophisticated system for both manual and automated rating.

3.8 Revised Rating Guide

The revised rating guide is the final iteration of our manual rating approaches. Due to the issues with a pure lexicon approach that motivated our switch to alternate methods of automation, we were no longer tied to a lexicon-based rating system. This allowed us to create a rating scale that better accounted for intensity of rhetoric while still incorporating the categories reflecting extremist narratives we had developed in our lexicon approach. We drew from our experience reading the comments to map the former lexicon categories onto the new scale. For example, we viewed calls to violent action as the most indicative of extremist ideology, as the legitimization

of violence is a key feature of extremist ideology.

We used a scale from 0 to 8, with a rating of 0 indicating no apparent extremist rhetoric and a rating of 8 indicating high presence of extremist rhetoric, including calls to violent action. Within the scale, we also had five levels, where if certain delineated features were present, the comment would be rated within that level at the lowest. Each level that indicated any amount of potential extremist rhetoric had two scale degrees (1-2, 3-4, 5-6, 7-8). This allowed raters some amount of flexibility and judgement within the levels without having too significant an impact on reliability. Through our past approaches, we had observed that some amount of subjective judgement was necessary, as one example of a feature of extremist ideas might be much more intense than another. As an example, while the comments "They just don't get us" and "We need to band together against those normies" both reflect us vs. them language, the latter sentiment is more intense and should be counted as such. However, we also wanted to limit the difference to one point on the scale to help avoid the inter-rater reliability problems we discovered with the categorical system.

We also decided to rate only on the explicit text stated, as we knew that our automation would not be able to read subtext. This included any text, including the text of links and text quoted from other comments, but not any content that could not be displayed in text such as images, videos, or webpages that a comment linked to.

The following section contains a more detailed overview of our rating criteria as outlined above. Please be warned that the remainder of this section consists of content that expresses incel-typical expressions of those rating criteria, including but not limited to swears, sexualized language, and suicidal sentiments. All examples are representative reconstructions of real comments,

Table 3.2: Revised Rating Guide

Level	Features Present
0	None
1-2	Name-calling, swearing, insults, non-bigoted insider language, self-directed insults
3-4	Bigoted language, us vs. them language, alarming statements that fall short of violent ideation or severe hatred, high-level insider language, insider language used as a directed insult
5-6	Severe hatred, slurs, strong us vs them language, strong problem with the world, suicidal ideation, violent ideation, admiration of violence
7-8	Calls to violent action, intent to commit violent action, expressed desire to perpetrate violent action

which will not be used for privacy reasons.[1]

0: Comments rated 0 have none of the features delineated in higher levels and do not seem to have any content related to extremist ideology.

Example:

- *"This is a woman"* (0)

1-2: Comments rated 1 or 2 have at least one of the following features: non-bigoted name-calling, swearing, insults, and minor uses of insider language. In the case of incels, minor uses of insider language would include words such as "incel" or "chad," which are specific

[1]For definitions of basic incel terminology, see `https://www.timsquirrell.com/blog/2018/5/30/a-definitive-guide-to-incels-part-two-the-blackpill-and-vocabulary`. For any additional definitions needed, reference `https://incels.wiki/w/Incel_Glossary`, which, while more thorough, is written by incels for incels, and as such should be treated as a primary source

to the community but are referenced and used outside of it. One way to differentiate between comments rated 0 and 1 are whether it would be uncomfortable to hear in a public setting. Many internet-typical comments fall under this section, and while we wouldn't characterize them as spreading extremist ideology, these comments contain more features of extremist ideology than the comments rated 0. Examples:

- *"Shut the fuck up."* (1)
- *"I'm an incel and proud of it"* (2)

3-4: Comments rated 3 or 4 have at least one of the following features: bigoted language, us vs. them language, alarming statements that fall short of violent ideation or severe hatred, high-level insider language, and insider language used as a directed insult. Us vs them language that falls into this level includes relatively mild blanket statements about the outgroup. In the case of incels, high-level uses of insider language would include words such as "mogged," "LDAR," or "ER"—language that is not typically known or used outside of the incel community. Examples:

- *"Why would an incel go to the gym just to get mogged by chads with lucky genetics?"* (3)
- *"Oh, you think I should 'just get on tinder'? Are you a white knight cuck or do you just want me to kms. Fucking normies."* (4)

5-6: Comments rated 5 or 6 have at least one of the following features: severe hatred, slurs, strong us vs. them language, strong problem with the world statements, suicidal ideation, violent ideation, and admiration of violence. For the purposes of this level, ideation is

not just mentioning violence or suicide. The comment counts as suicidal ideation if it directly says or strongly implies that someone wants to die or kill themself and it counts as violent ideation if it directly says or strongly implies that they approve of violence or violent action. Examples of violent ideation and violence admiration include praising mass murderers (e.g. "Saint ER") or espousing support for violence without directly stating that they would perpetrate the violence. Examples:

- *"Roasties know that they'll hit the wall which is why they spend their 20s riding the CC and then try to lock down a betabux that they can tolerate monthly sex with."* (5)

- *"Watching foids taunt men with their tight leggings makes me understand saint ER-cel."* (6)

7-8: Comments rated 7 or 8 have at least one of the following features: calls to violent action, intent to commit violent action, and expressed desire to perpetrate violent action. These differ from ideation and admiration in that it puts the author in the role of the perpetrator. Examples:

- *"Well I'm going to turn 30 in 8 hours and if I haven't had sex by then (kek) I'm just going to rope. Thanks for the laughs, you guys were the only thing making my inceldom bearable all these years."* (7)

- *"Like 80 percent of men are incels; its time to rise up and force the government to exterminate cumskin foids and provide every incel with an asian virgin for them to marry."* (8)

These comments display severe racism and hatred of women combined with an explicit call for violence.

3.9 Manual Use of Revised Rating Guide

Using the revised rating guide, we selected sets of comments and split the data so that every comment was rated by two people. This was separated into four rounds, so that we could test and modify the rating guide based on any points of confusion encountered while rating.

Throughout each round, every comment was rated by exactly two people. This was to ensure that we would have some way to check our classifications while also maximizing the amount of labeled data available for training. Figure 3.1 contains the distributions of the difference between a comment's two extremism ratings, separated by labeling round. These data exclude irrelevant comments, such as those posted by bots or comments which are unrelated to any aforementioned features of extremist narratives. After the first round, we also balanced the comment distribution so that every pair of raters overlapped on the same number of comments.

As the larger tail in the first round of Figure 3.1 shows, there was initial confusion using the revised rating guide. This round also had other miscommunication difficulties related to who was meant to rate which comments. This discrepancy was largely removed by the later rounds, with a large majority of comments having identical or almost-identical ratings. Later rounds also included an option for raters to flag comments for review, which we used to update and clarify the rating guide. For analysis of inter-rater reliability of these rounds, Krippendorff's α is more appropriate than Fleiss' κ because the ratings based on the revised rating guide are ordinal—a higher rating implies more extreme content—and each comment was only rated by

two people, causing "missing" data that would impede other reliability measures [22]. An α score of 1 indicates perfect agreement, while an α of 0 indicates agreement no better than pure chance. Table 3.3 contains the α value for each round of rating using the revised rating guide. There was a great improvement in reliability between rounds 1 and 2, with slight decreases in reliability afterwards in rounds 3 and 4. Krippendorff recommends relying only on variables with $\alpha \geq 0.800$, and drawing tenuous conclusions from variables with $\alpha < 0.800$ only when $\alpha \geq 0.667$ [22]. This implies that our ratings, at best, can only offer tenuous conclusions, though the revised rating guide offers a significant improvement over the reliability of our previous rating method. While there was still some disagreement between manual ratings, we decided that this level of agreement was sufficient for moving forward into automated detection.

Table 3.3: Revised Rating Guide Inter-rater Reliability

Labeling Round	Krippendorff's α
1	0.511
2	0.741
3	0.711
4	0.675

Chapter 4: Automated Extremism Detection

In order to identify the extreme comments within our dataset (consisting of about 1,000,000 samples), we needed an automated system which would faithfully reflect our definition of extremism as described in prior sections. This is effectively a binary classification natural language processing task. We implemented three different approaches with increasing levels of sophistication before

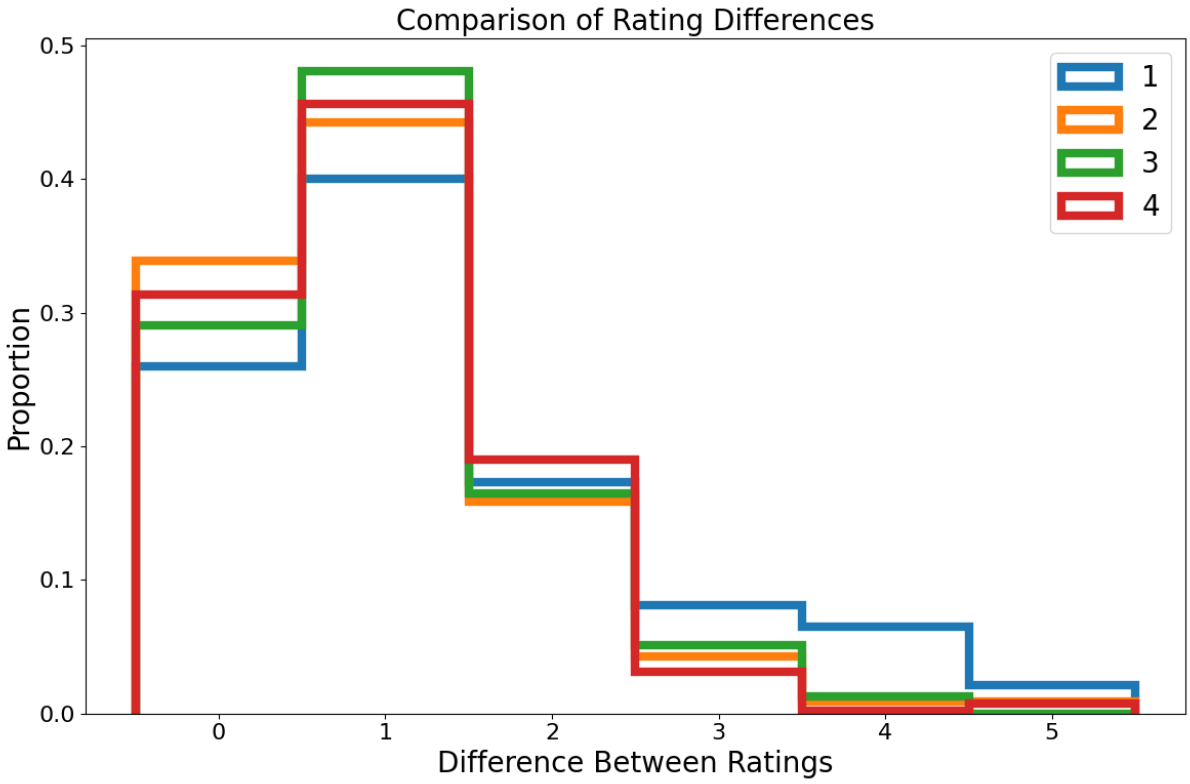

Figure 3.1: The distribution of differences in ratings for every manually rated comment.

obtaining a model with sufficient accuracy:

- Weighted Lexicon Scoring (hand-tuned)
- Logistic Regression on TF-IDF Embeddings
- BERT: Transformer-Based Language Model

The latter two fall under machine learning. These more advanced techniques come at the cost of interpretability but provide significantly more discriminate power. BERT is considered state-of-the-art [23] and provided us with the best performance, demonstrating accuracy in the 95% range. The performance of all three is summarized in Figure 4.1 by their receiver operating characteristics. Additionally, confusion matrices are shown for each approach with the threshold chosen to maximize weighted-accuracy in the following sections.

4.1 Training Data

To serve as training data, the team manually rated ∼2500 comments from `r/Incels`. This process is described in more detail in sections 3.7 and 3.8. For cases in which a binary label was required, a score of 3.5 out of 8 was used as the threshold, resulting in a 10% positivity rate.

The data was split 70-10-20 into training, validation, and testing sets. Care was taken to ensure that any comment authors who appeared multiple times in our manually labelled data were fully isolated in one of these splits. This avoids the possibility of the model learning to identify particular authors who happen to write extreme content as opposed to extreme content itself. However, such instance of repeated authors were rare.

The random sample of comments rarely included non-English posts. These were not rated

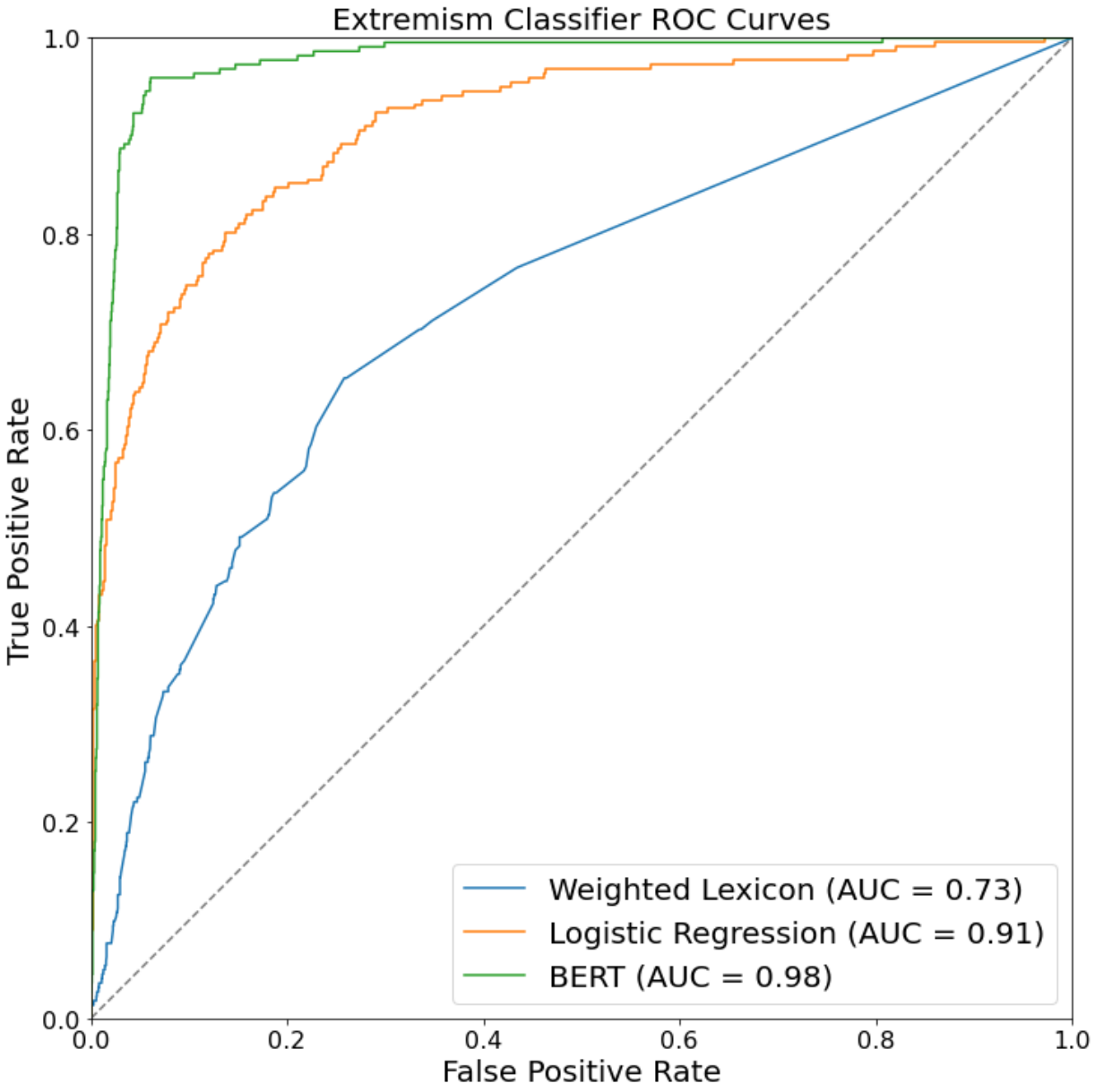

Figure 4.1: The receiver operating characteristics for all three classifiers.

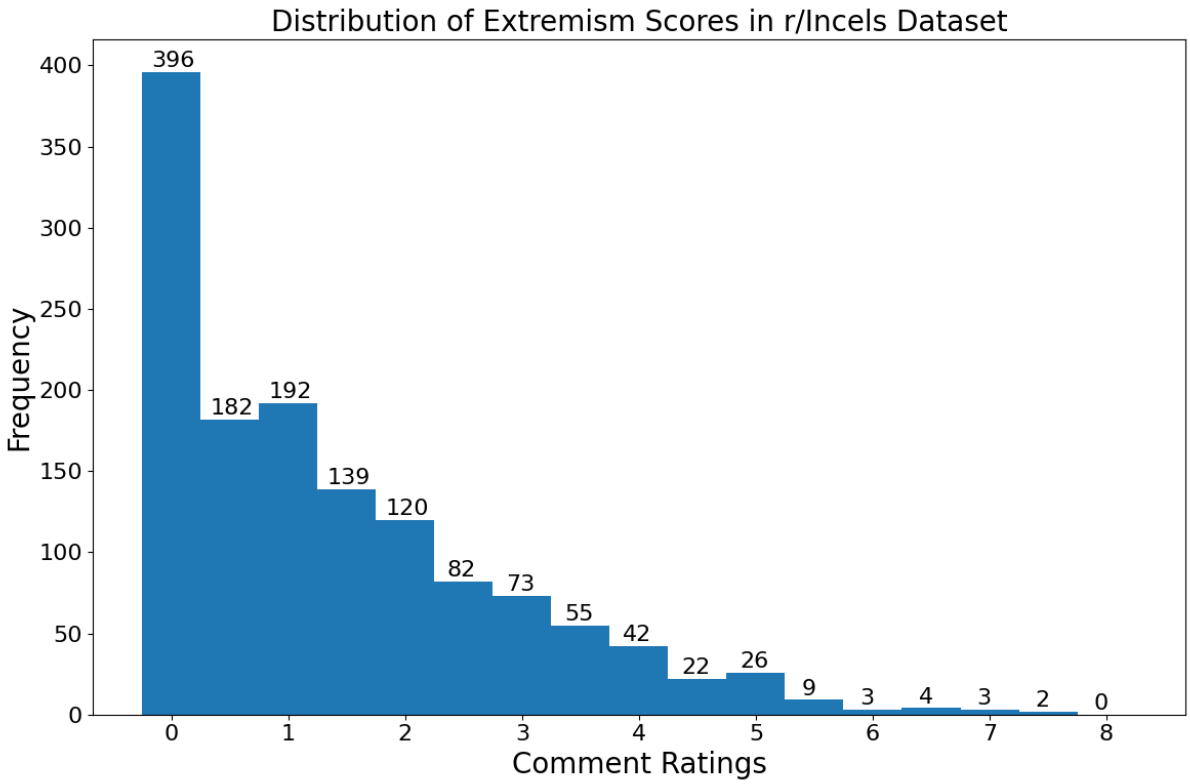

Figure 4.2: The distribution of manual ratings over the full `r/Incels` dataset. Very few comments received a score ≥ 5.

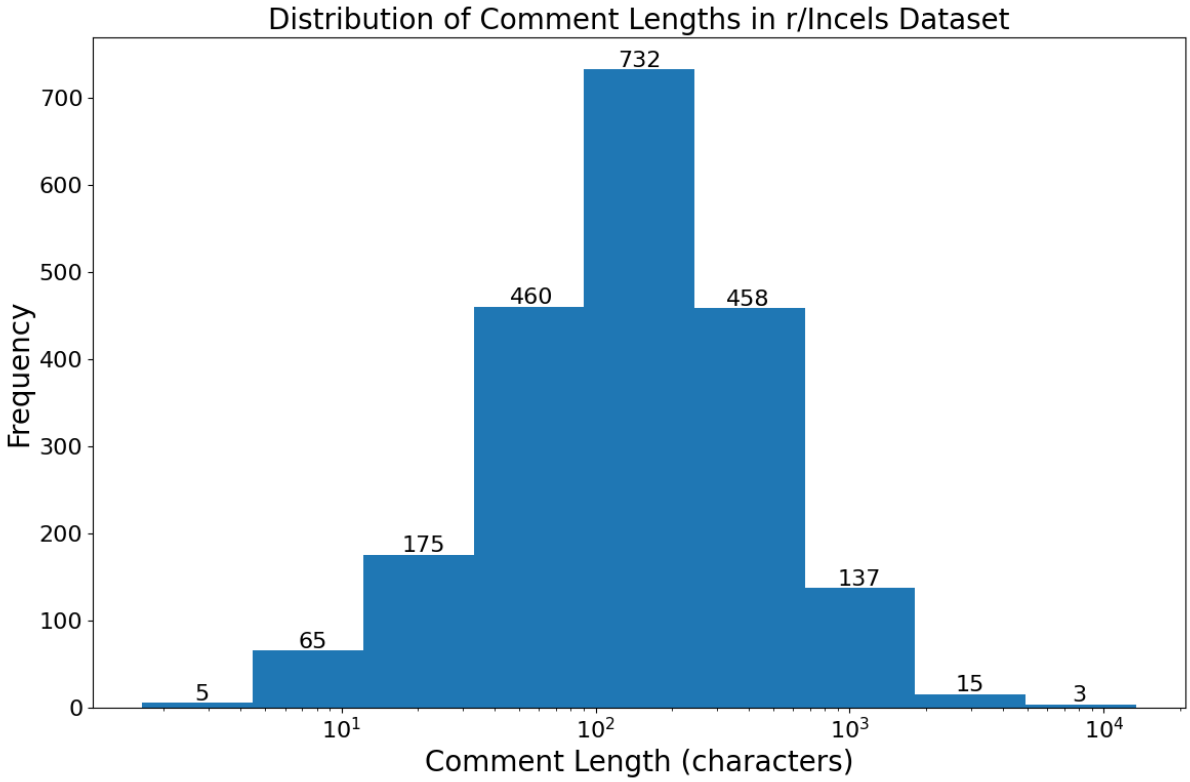

Figure 4.3: The distribution of comment lengths over the full `r/Incels` dataset. There is a long tail and a small number of outliers. Very few comments are more than 1000 characters long with the median length being only 93 characters.

or included in the training set. Because of this, our model would not be effective in identifying extremism in communities which do not speak English. BERT, however, can be trained on a number of non-English languages, which would allow for future training sets to be composed of non-English comments.

4.2 Weighted Lexicon Scoring

The weighted lexicon approach identifies extreme content by operating at the word level. The team manually compiled a list of terms which we subjectively associated with extremism. Each term fell into one of seven categories. Each category was assigned a weight based on their intensity and contribution to an extremist narrative:

Table 4.1: Lexicon Categories and Weights

Lexicon Category	Example	Weight
profanity	`bitch`	1
insider language	`volcel`	2
problem with the world statements	`injustice`	2
harassment	`fatass`	3
us vs. them rhetoric	`they`	5
hate speech	`<ethnic-slur>`	6
calls to violence	`euthanize`	8

For example, profanity, which is common on the internet and not always indicative of extremist ideology, was weighted much lower than violent language, which is much less common and frequently indicative of extremism. The numerical weights were chosen largely subjectively based on the team's past exposure to the data, with the notion of density in mind. Suppose a comment containing at least 10% violent language is considered extreme. Then, a comment must contain at least 80% profanity, or similarly 40% insider language, to be similarly classified. A comment's score is computed as follows: first, a weighted sum of the activation of the categories

is calculated (i.e. category weight $\times$ number of matched terms in the corresponding lexicon, summed over the categories). Then, the sum is normalized by the sigmoid function:

$$f(x) = \frac{1}{1 + e^{-x}}$$

This maps $[0, \infty)$ to $[0, 1]$.

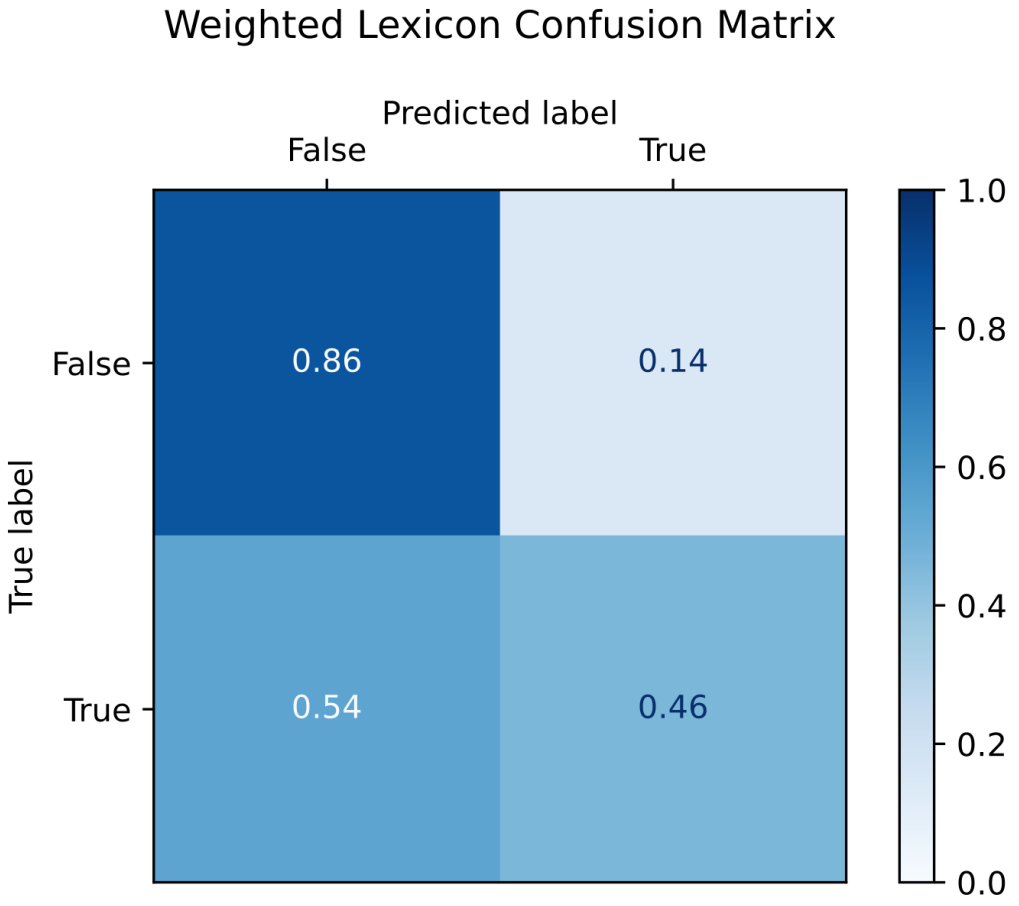

Figure 4.4: Weighted Lexicon System confusion matrix (normalized by row). Classification threshold chosen to maximize weighted accuracy. Recall is poor.

While this approach is easily implementable and modular, it struggles to address words used in common parlance which were particularly indicative of extremist ideology in the context of `r/Incels`. For example, words for genders such as "man" and "woman" did not strictly

fall into any of our distinguished lexicon categories, but generalizations of these groups were common in more extreme comments.

Additionally, some phrases or combinations of words indicated extremist ideas while the presence of any word individually would not. We augmented our lexicon with some of the most common multi-word phrases. Also, the presence of matched terms in multiple lexicons was typically indicative of a more extreme comment, and certain pairs of categories had a strong tendency toward extremist ideology. Thus, we added a layer to the score calculation: the weighted sum was amplified according to the number of distinct lexicons represented, and further amplified if certain pairs of coupled categories were present (e.g. *profanity* and *harassment*, *problem with the world statements*, and *calls to violence*).

Nevertheless, the lexicon approach was insufficient. Because the presence of sensitive language does not always indicate agreement with or intentional discussion of extremist ideas, we considered implementing sentiment analysis. This would assess the author's attitude toward the ideas delineated in the comment as either positive or negative. Some of the words classified in the *calls to violence* category could be used in innocuous contexts, resulting in phrases like "You don't want to feel burned" receiving much higher extremism scores than intended. These issues, combined with an inability to distinguish severity of words within the same category, meant that the system did not hold much room for nuance.

Finally, while we gathered the lists of words in each category from thorough sources such as Hatebase.org, new words would consistently need to be added as we reviewed comments which should be considered extreme but were not appropriately labeled by our method. It became clear that more sophisticated approaches were required in order to encapsulate our nuanced definition of "extreme."

4.3 TF-IDF and Logistic Regression

Logistic regression is a standard technique for binary classification and has been previously used in rudimentary forms of sentiment analysis [24]. It operates by assigning positive and negative weights to certain words, which it learns from training data. In order to implement logistic regression, we needed a way of encoding comments as vectors. This is a standard procedure in natural language processing, with many pre-built solutions. We opted to use TF-IDF, which allocates each dimension of the vector space to one word in the corpus vocabulary.

For a given comment, TF-IDF assigns values to these dimensions by computing the normalized frequency of the corresponding word and then transforming the value based on its prevalence in the entire corpus [24]. We compute a vector representation for each comment in the dataset and then train the logistic regression model using standard gradient descent and binary cross-entropy loss. The results of running the classifier on the testing dataset, shown in 4.5, are much improved from the weighted lexicon approach, but still fall very short of human performance.

This approach addresses many of the issues of the lexicon system since it is able to learn a different weight for every term in the entire corpus vocabulary. Some of the terms which received the highest weights were “males,” “incel,” “dead,” “normies,” and “men,” which provided a reality check that the model was broadly capturing the desired trends. However, the model’s ability to reliably detect extreme comments was deemed inadequate. From a subjective analysis of the ratings produced by logistic regression, it became clear that we need a language model capable of detecting relationships between words.

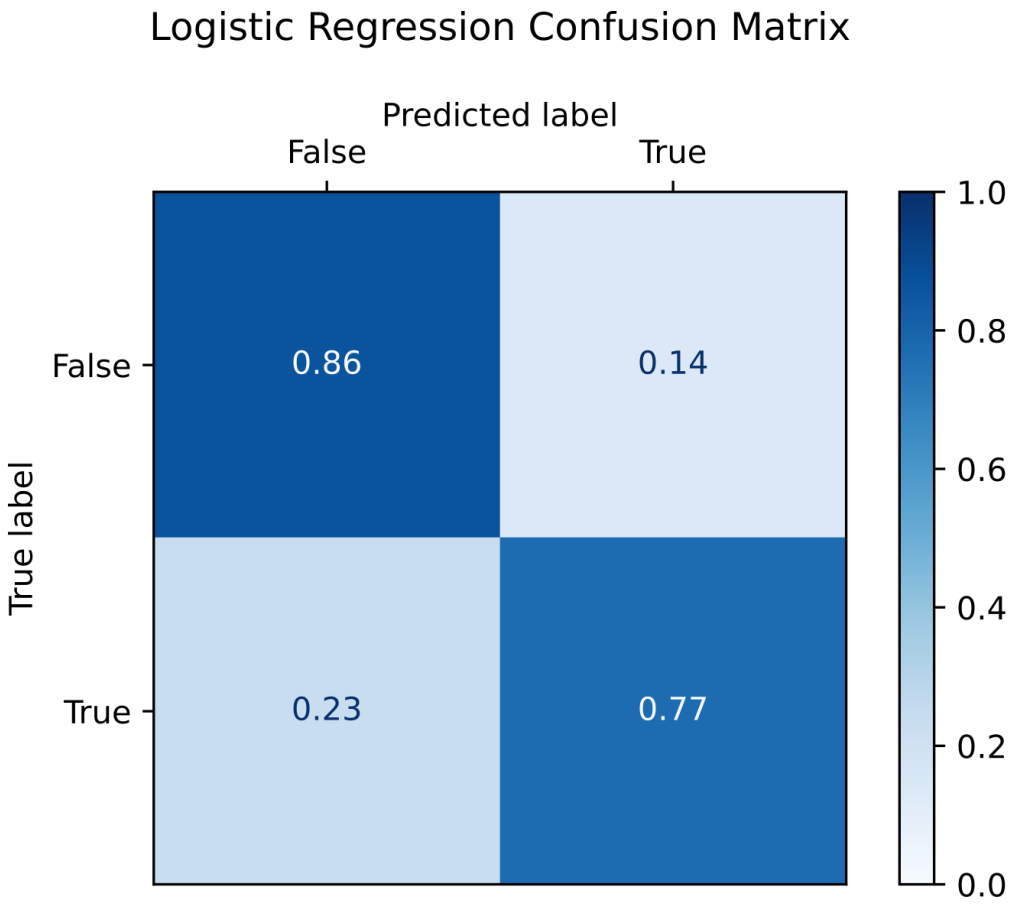

Figure 4.5: Logistic Regression classifier confusion matrix (normalized by row). Recall is significantly improved from lexicon approach, but the false-negative rate is still unacceptably high.

4.4 BERT Language Model

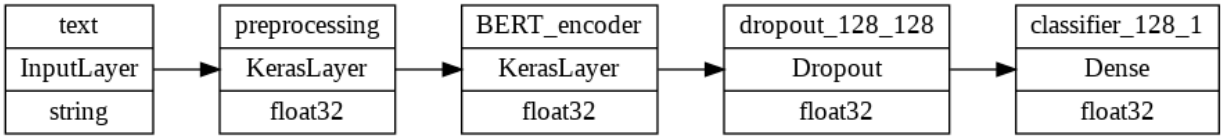

Figure 4.6: Model architecture of BERT-based classifier. Text is first cleaned, tokenized, and vectorized in the preprocessing layer. Then BERT produces an encoding which is passed to a dense classifier layer with 128 nodes. Total parameters: 4,386,050

BERT is a transformer-based language model developed by Google. It was first published in 2018 and is considered state-of-the-art[23]. Pre-trained models are freely available which have been trained on huge data sets including all of English Wikipedia. We selected one such model, specifically `bert_en_uncased_L-2_H-128_A-2`, which is considered a small version of BERT despite having over 4,000,000 trainable parameters. We then fine-tuned the model on our labelled data.

Figure 4.6 shows the full classifier architecture, including a fully connected final layer with 128 nodes. The first step in the architecture, preprocessing, works by coercing the text to lowercase, removing accents and other non-standard characters, and tokenizing. It uses a vocabulary collected from the pre-training stage consisting of all of Enlgish Wikipedia and BooksCorpus. This includes virtually all of the terminology from our data set, including incel-specific terminology. Fewer than 1% of terms were mapped to the `[UNK]` token. Training the model took about 1 CPU-hour, and the results are very promising.

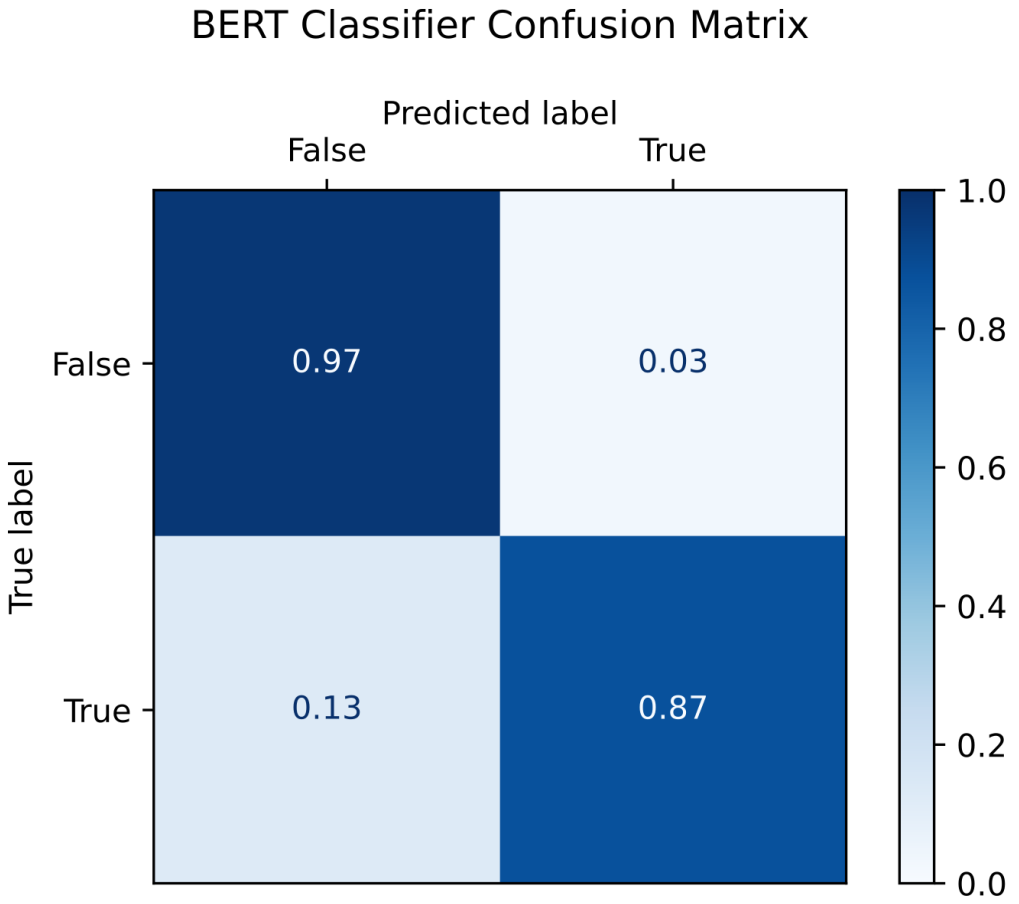

Figure 4.7: BERT classifier confusion matrix (normalized by row). Precision and recall are both greatly improved from prior approaches.

4.4.1 Measuring Performance of the BERT Model

When performing our manual classification, there was a significant difference between rounds 1-3 and round 4; while the first few rounds were randomly sampled from a database of all comments on the subreddit, round 4 was taken specifically from comments that the BERT model judged to be on the border of not-extreme and extreme. The purpose of this was twofold: it allowed us a way to test the model's performance and provided the best set of additional training data with which to refine the model.

Figure 4.8 demonstrates another angle of the BERT model's effectiveness. There is a clear qualitative difference between the comments that the model was unsure of and the overall distribution of comments in the subreddit. Of particular note is the fact that this data has a peak at a rating of 3, which is right around the level that we selected as a cutoff between extreme and not-extreme. Also interesting is the sudden drop-off at a rating of 6, as the model was much more sure that these comments were extreme.

The fact that these comments range in large numbers from ratings of 0 up to 5 is not perfect, but the results are still promising and would likely be significantly improved if we were to include this data to train the model. Due to time concerns, we are not able to perform an additional test of the data that would result, so we decided to leave the model as-is while we have a measurement of its effectiveness. Implementing and testing the additional training data from manual classification

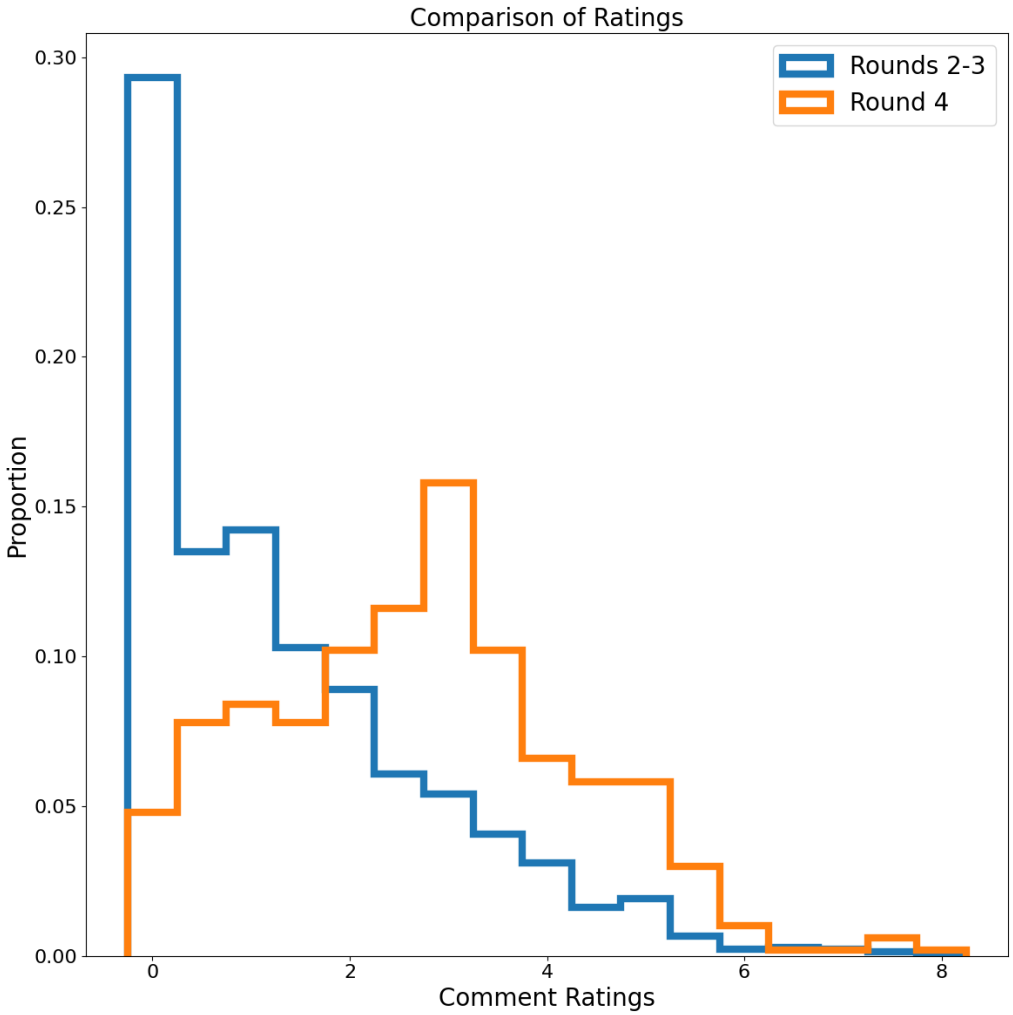

Figure 4.8: A comparison of the ratings for randomly selected comments (Rounds 2-3) and the comments that the BERT Model was least sure about (Round 4).

round 4 presents a clear way to extend our results in the future.

Chapter 5: Mathematical Model of Extremism Propagation

Our model draws inspiration from the work of Jabin and Motsch [17] on opinion dynamics and consensus formation, which uses the consensus model of Krause, Motsch and Tadmor [25] [26] [27]. While both this model and ours are agent-based, their interpretations differ in some key ways. Most notable among these is that our model should not be interpreted as predictive for individual agents; it is only intended to be predictive in aggregate. Ostensibly, an agent based model attempts to predict the way individual agents will change in the future, but there are too many invisible variables acting on each user in this context for any such individual prediction to be reliable. In aggregate, however, the results of the model can be used to predict the behaviour and makeup of the subreddit as a whole. This technique of adapting the agent-based approach has precedent, and is used in some models of population growth [28] [29], among other things.

For practicality of computation, our model is discrete in time. The fundamental units of our model are agents that vary in opinion over each discrete time-block, and comments that occur during single time-blocks and have fixed extremism values. We model an agent as interacting with each comment in a time-block by observing their content and reacting to the extremism present. This provokes a change in opinion proportional to the agent's engagement with the subreddit and the comment's visibility to users. Interactions are mediated by $\phi(O_i, E_j)$, a function of the opinion of an agent and the extremism of the comment they are observing, which

determines the strength of the agent's reaction. One notable difference between our use of ϕ and that of Jabin and Motsch is that we allow ϕ to be negative; that is, we consider it possible for an agent to observe a comment and agree less with it than they would have if they hadn't seen it. This behavior is not desirable in the consensus-formation setting, but here we do not expect opinion convergence.

We discretize our model by breaking time up into evenly sized blocks during which comments are generated and observed. We chose time-blocks of 28 days so as to smooth out errors created by discretization while ensuring that our data would not be too reduced. During each time-block, the population of agents changes, with some selection of agents leaving the environment and some new agents being introduced. Agent entry and exit is dependent on a number of factors, including agent opinion, extremism of the subreddit, and quantity of active agents. Those agents which do not leave have their extremism modified by each comment belonging to the current time-block, mediated by ϕ, visibility, and engagement as described above. We normalize total change in extremism for an agent by the total visibility of the comments they view, limiting the total attention bandwidth the agent can devote to the subreddit.

5.1 Agent and Comment Qualities

We index the set of all agents by $i \in I$, and the set of all comments by $j \in J$. We index the set of time-blocks by t, noting that each t corresponds to a fixed interval of time, not a specific point. Under this notation, each agent corresponds to one Reddit user who actively posts in the set of time-blocks T_i. As comments are relevant for a few days at most, which is brief relative to our time-blocks, they are modeled as existing for only one period t_j.

Formally, agents are defined by three qualities. The first two, **Opinion** and **Engagement**, are functions of time, where $O_{i,t}$ denotes the opinion of an agent at the end of time-block t, and $N_{i,t}$ likewise for engagement. The third quality is the **period of activity**, which is the set of time-blocks T_i in which the agent posts at least one comment. We define

$$A_i := (O_{i,t}, N_{i,t}, T_i,)$$

$$\mathcal{A} := \{A_i\}_{i \in I}$$

$$\mathcal{A}[t] := \{A_i | t \in T_i\}_{i \in I}$$

where $\mathcal{A}$ is the set of all agents, and $\mathcal{A}[t]$ is the set of all agents active in time-block t.

Conceptually, opinion is a measure of each agent's internal extremism, and modeling global changes in $\{O_{i,t}\}_{i \in I}$ across time is the primary focus of our model. Engagement is a measure of how frequently an agent interacts with the subreddit, and it determines how rapidly an agent changes their $O_{i,t}$. While nominally varying across time, in our model, $E_{i,t}$ is treated as constant across time for each agent.

Similarly, each comment is defined by four qualities: **Extremism**, denoted E_j, **Visibility**, denoted V_j, **Author** A_j, and **time-block** t_j. Formally, following the previous notation,

$$C_j := (E_j, V_j, A_j, t_j)$$

$$\mathcal{C} := \{C_j\}_{j \in J}$$

$$\mathcal{C}[t] := \{C_j : t_j = t\}_{j \in J}$$

$$\mathcal{C}[A] := \{C_j : A_j = A\}_{j \in J}$$

Here, extremism denotes the extremism content of that particular comment, as evaluated by the BERT classifier. Visibility denotes the relative likelihood that a comment is read, and will be modeled as a function of the upvotes a comment receives. Author denotes the agent which produced the comment, and time-block is the block during which the comment was produced.

5.2 Theoretical Behavior

Conceptually, the extremism of the subreddit as a whole is dependent on two things. First, at the agent-level, each individual person's opinion may change based on their interactions with the subreddit. Second, at the population level, people may join or leave the subreddit, potentially based on their own opinion and the subreddit's overall extremism level. To account for this, our model incorporates both agent-level and population level changes.

At the agent level, each agent is predicted to adjust their opinion each timestep based on their engagement and the comments they observed, according to the following *modified consensus model*.

$$\Delta O_{i,t} = N_i \frac{\sum_{C_j \in \mathcal{C}[t]} \phi(O_{i,t}, E_j)(O_{i,t} - E_j)V_j}{\sum_{C_j \in \mathcal{C}[t]} V_j} \tag{5.1}$$

This model describes how the interaction between agents and comments occurs. Conceptually, each comment seen has the capacity to influence an agent's opinion. The degree to which seeing a comment influences each agent, however, varies. An agent's receptiveness to a comment is a function of their current opinion and the comment's extremism, which we denote by $\phi(O_{i,t}, E_j)$.

Mathematically, when an agent of opinion $O_{i,t}$ views a comment of extremism E_j, we consider their opinion to change by $\phi(O_{i,t}, E_j)(O_{i,t} - E_j)$. It is important to note that ϕ may take negative values, representing the scenario in which a sufficiently extreme comment repulses more moderate readers, thereby pushing them away from extremism. This is in contrast with the model proposed by Jabin and Motsch [17], in which ϕ is strictly positive.

Ideally, we would know the set of comments each agent has seen, and compute their change in extremism accordingly. However, while Reddit does have a system of 'upvotes,' which tracks how many users liked a comment, it does not provide data on *who* upvoted each comment. Thus, we instead estimate the effect of reading comments probabilistically. More specifically, using the fact that V_j is the relative likelihood that a comment is seen, we can estimate the average effect of reading one comment as:

$$\frac{\sum_{C_j \in \mathcal{C}[t]} \phi(O_{i,t}, E_j)(O_{i,t} - E_j)V_j}{\sum_{C_j \in \mathcal{C}[t]} V_j}$$

Finally, scaling this by each agent's engagement accounts for the relative activity of agents on Reddit. This gives us our final model form seen in equation 5.1.

At the population level, we will use more standard probabilities of entry and exit, which we will estimate using a logistic regression. Predictions for the next time-period are then generated probabilistically using the results. For entry, we cannot do that, because we cannot observe the population of people *potentially* entering the subreddit. We thus instead model the number of people entering the subreddit each day as a global function of subreddit attributes, which will be obtained through linear regression.

5.3 Estimating variables

As discussed in introduction, our database consists of all comments from the subreddit `r/Incels` from the period of May 2016 to early November 2017, when the subreddit was banned. For each comment, we have:

1. The content of the comment,
2. The username of the user who posted the comment,
3. The final 'score' of the comment, which is the total number of 'upvotes' minus the total number of 'downvotes' the comment received over its lifetime, and
4. The time the comment was posted.

Using this information, we derive approximations for our variables $E_j, V_j, O_{i,t}$, and $N_{i,t}$. E_j is determined entirely by the content of the comment, specifically by using the machine-learning model BERT, as described in Section 4.4. For our purposes here, E_j is a binary variable, where score of 4 or higher under our rating guide corresponds to $E_j = 1$, while a score of 3 or lower corresponds to $E_j = 0$. This transformation is due to the fact that BERT is a binary classifier, and so can only return confidence values for whether a comment is above or below a threshold. Future work will entail using BERT to classify comments into our existing 0 to 8 rating system.

We approximate V_j using a much simpler system. On Reddit, each post has beneath it a list of comments, each with a score determined by subtracting total downvotes from total upvotes. By default this list is sorted from top-to-bottom by a metric favoring comments with higher score. Thus, at each time, the visibility of a comment is conceptually an increasing function of its

score. While we cannot directly obtain score at each time, we can expect that on average, higher final score is highly correlated with lifetime visibility, especially over longer time-blocks. Thus, we model V_j as an function of a comment's final score, with the functional form determined by testing various functions on our training data. Specifically, preliminary testing shows that $V_j = score$ is both the simplest and best performing function, although this result will be revisited as a part of future work.

Approximating $O_{i,t}$ is somewhat more challenging. In general, the only information we have about an agent's opinion is from the extremist content they post. In trying to model the spread of extremism on social media, however, what we hope and seek to predict is in fact the extremist content posted in the future. Thus, for our purposes, is it reasonable to treat $O_{i,t}$ as a measure of the average extremism of comments made by agent A_i.

More formally, we assume that an agent's opinion determines the average extremism of their comments in the next time-block, where opinion is a time-discounted average of past comment extremism, according to the following formula:

$$P_{i,t+1} = \frac{\sum_{C_j \in \mathcal{C}[A_i]} E_j \beta^{t-t_j}}{\sum_{C_j \in \mathcal{C}[A_i]} \beta^{t-t_j}} \tag{5.2}$$

where β is a time-discounting factor. To obtain β, we simply apply a mean squared error function to compare our predictions to actual average extremism and test between $0 < \beta \leq 1$ to identify the β which minimizes our error across our dataset. The result of our search determined that $\beta \approx \frac{125}{128}$ would be the best choice.

To approximate engagement, we obtain an average ratio between the agent's actual change in opinion over time and their "expected" change, i.e. a move towards the mean extremism of the

subreddit as a whole. This is given by the formula

$$N_{i,t} = \frac{1}{|T_i'|} \frac{\sum_{t \in T_i'} O_{i,t+1} - O_{i,t}}{\sum_{t \in T_i'} (\sum_{C_j \in \mathcal{C}[t]} E_j v_j) - O_{i,t}} \tag{5.3}$$

where T_i' is the set of time-blocks t during which agent A_i is active during both t and $t+1$. We note that $N_{i,t}$, while theoretically varying in time, is treated as a constant within our model. This is because an excess of degrees of freedom for an individual agent across time would allow for extreme overfitting and limit generalizability.

Finally, in order to approximate the interaction function ϕ, we rely on more sophisticated gradient descent methods, which are described in detail in Section 5.4.

5.4 Obtaining ϕ - Gradient Descent

The most critical component to the function of the model is the interaction function ϕ, which describes the direction and strength with which an agent of opinion $O_{i,t}$ is influenced by a comment of extremism E_j. We model ϕ as piecewise constant, which simplifies and expedites the computation used to obtain it while preserving its descriptive power. We identify each constant piece of ϕ as a variable $\phi_{i,j}$, which corresponds to $\phi(O_{i,t}, E_j)$ for a cross section of opinions and extremism values. We cannot obtain these $\phi_{i,j}$ directly. Instead, we perform gradient descent to obtain these parameters.

To perform gradient descent, we examine time-block t and iterate the following process. Take the set of agents $\mathcal{A}[t]$ who make some comment occurring in t. We avoid training on agents who have not posted any comments in t because doing so would introduce undue bias against predicting any change in an agent's opinion. We then calculate the gradient ∇MSE of our

mean-squared-error loss function

$$MSE = \frac{1}{|A_i|} \sum_{A_i \in \mathcal{A}[t]} (O_{i,t} - O_{i,t-1} - \Delta O_{i,t})^2 \tag{5.4}$$

with respect to our variables $\phi_{i,j}$. Here, $\Delta O_{i,t}$ is obtained by equation 5.1. We note that equation 5.4 is quadratic with respect to each parameter, so ∇MSE is trivial to calculate. Let $\vec{p}_n$ be the vector of our parameters at iteration n of gradient descent, where $\vec{p}_0$ is chosen uniformly randomly from its domain. We descend in the opposite direction of the gradient by updating our parameters according to

$$\vec{p}_{n+1} = \vec{p}_n - \gamma_n \nabla MSE$$

where γ is a scale factor. To improve convergence by decreasing our step size over time, we define our scaling factor at iteration i to be

$$\gamma_i = \gamma_0 \cdot \frac{1000}{1000 + i} \tag{5.5}$$

After many iterations of this process, the resultant parameter vector $\vec{p}$ selects values which have some predictive power in aggregate, which are then used in our model.

5.5 Obtaining Entry and Exit - Regression

While the interaction function ϕ governs how each agent changes their behavior, it does not capture the change in overall extremism due to changes in the overall population of agents. This is important in our period for two reasons. First, because ϕ governs the change in extremism after entry, the extremism of the first comment posted is not determined by our above model.

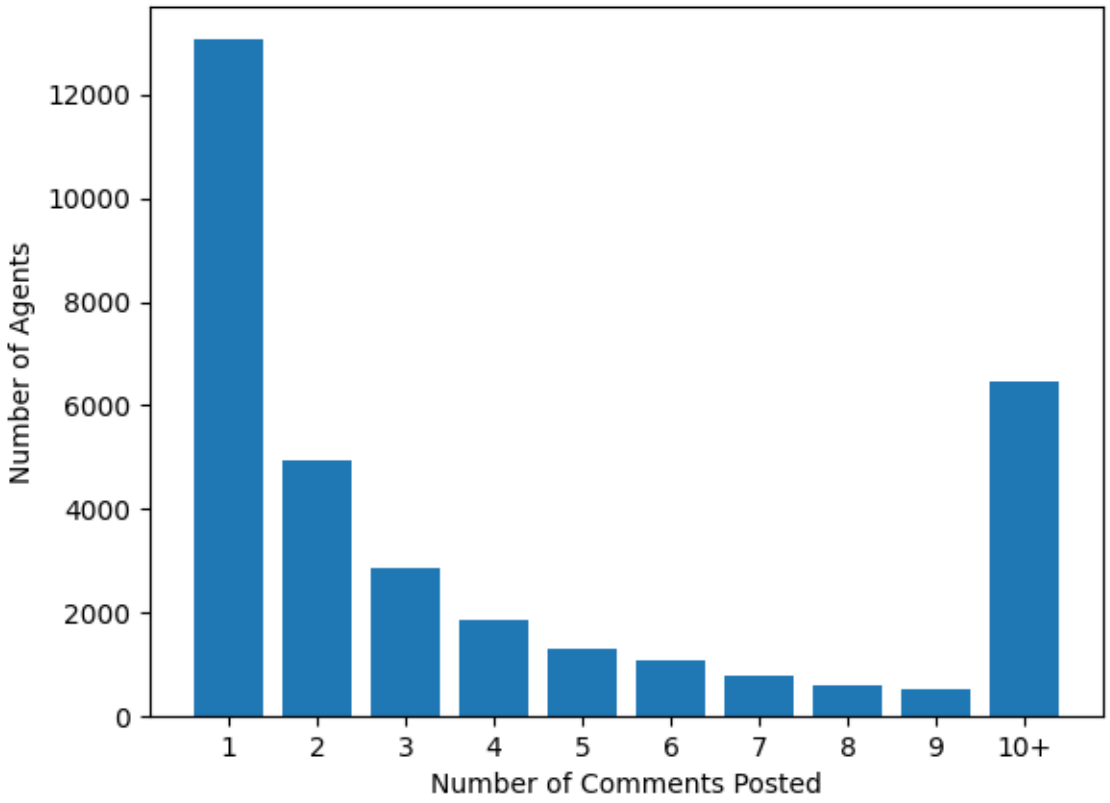

Figure 5.1: Distribution of total number of comments posted by each user.

Simultaneously, the majority of users on Reddit post relatively few comments. As seen in Figure 5.1, it is very common for an agent to post exactly one comment, and a comfortable majority post five or fewer. For these agents, their extremism when entering the subreddit remains a significant factor in determining their overall extremism. In particular, the more than 13,000 agents who only post once comprise over one third of all users, and are determined completely by their entry extremism.

Second, the subreddit experienced major growth between the years of 2016 to 2017, which is our period of study. As seen in Figure 5.2, the daily number posts during this period steadily rose until the subreddit's banning. This means that accounting for overall growth or decline is an

important component of making medium to long term predictions.

To begin accounting for entry and exit, we first observe two important qualities in Figure 5.3, which shows the total number of new users each day. First, we observe that behavior in the first few months of the subreddit is basically non-existent, which points to the potential of erratic behavior due to small sample size. When we plot the number of users who joined the subreddit relative to the population as a whole, as in Figure 5.4, we see this suspicion is verified. This makes intuitive sense, as during the initial months of 2016, which is when the data was taken, the subreddit was just beginning a period of sudden growth. Thus, the data concerning entry and exit from the initial periods are fundamentally different from the following data. To account for this in our regressions, we simply drop the first three months of data, to obtain a more behaviorally consistent sample.

Second, looking at Figure 5.2 as well, there are clear and distinct periods where almost no comments were posted and no new users joined, with the largest such gap at around the 430th day. This is because `r/Incels`, throughout its lifespan, occasionally went private or briefly self-regulated in the face of public scrutiny or scandal. In total, there are six such periods in our model, with four being short three to five day gaps. This posed a conceptual challenge, because our agent based model was based on 28-day time-blocks, whereas to properly account for these short gaps using dummy variables requires more granularity. Ultimately, we chose to analyze entry and exit on a day-by-day basis, and simply compute the total entry and exit over each 28-day block at the end. This allows us to accurately account for these gaps, but has the potential flaw of compounding error, as each day's entry/exit prediction depends in some part on the outcome of the last.

We also have to account for what we can observe about agents. Conceptually, an agent may

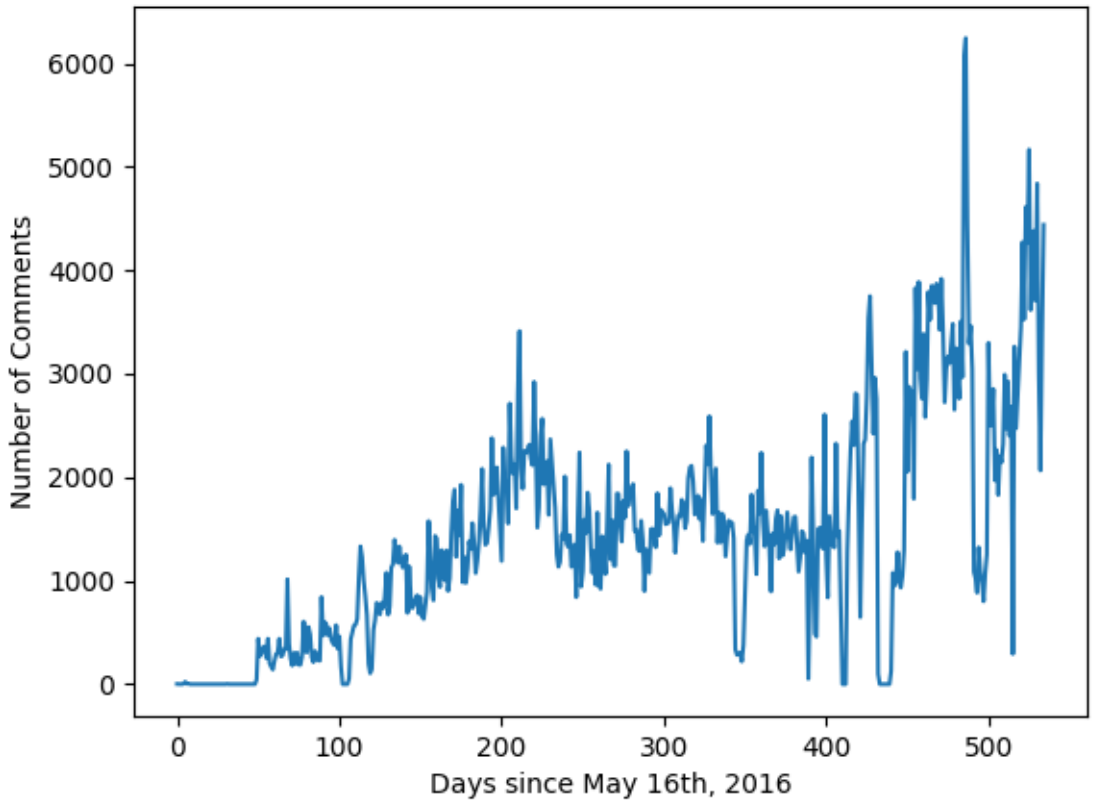

Figure 5.2: Number of New Comments by Day

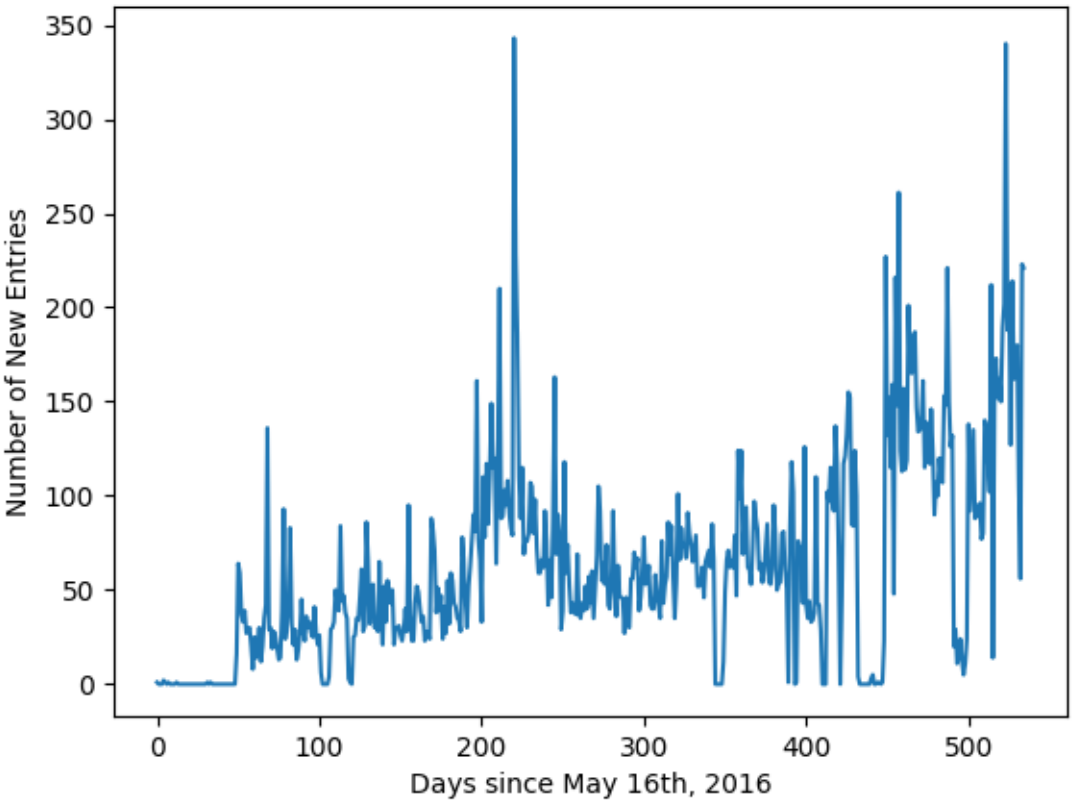

Figure 5.3: Number of New Entries by Day

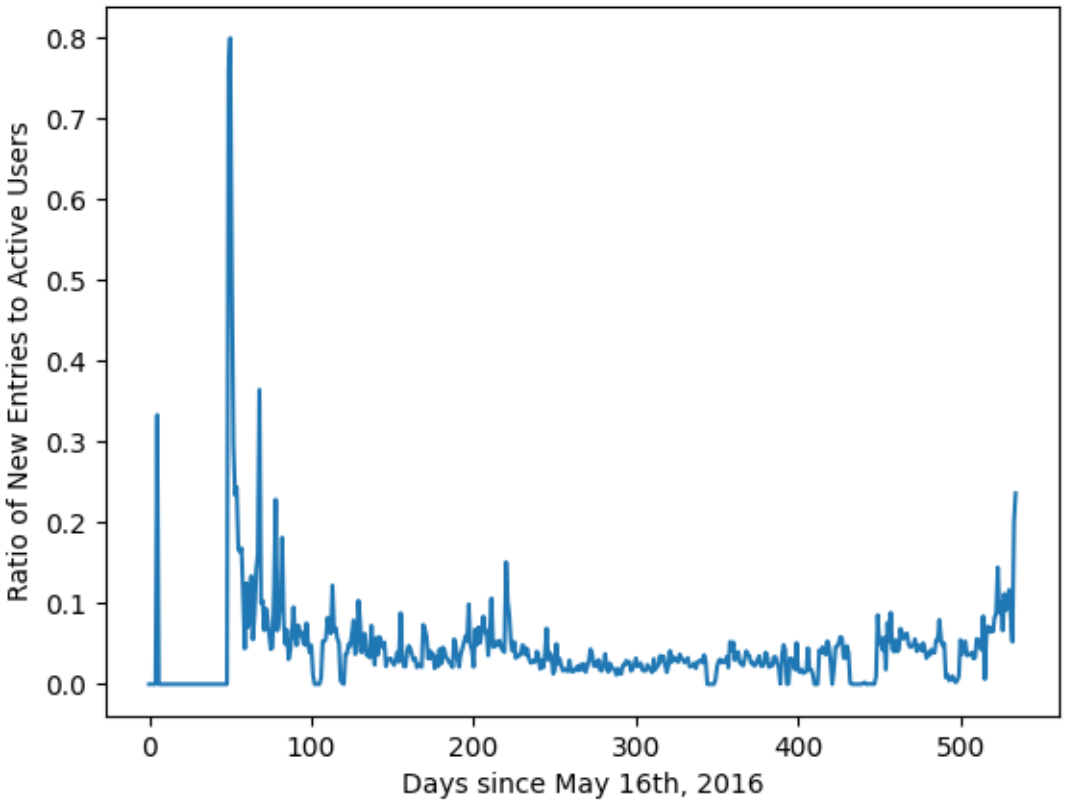

Figure 5.4: Ratio of New Entries to Active Users by Day

enter the subreddit with any arbitrary value of extremism in our range. However, we can only calculate agent opinion based on extremism of comments posted, and as seen in Figure 5.1, most agents post rarely or only once. This means that an agent's second comment, if it even exists, is generally a poor indicator of entry opinion. Moreover, any measure of entry extremism that tries to incorporate future comments will simply not apply to users who only post once, which is one of our principal concerns. Thus, instead of modeling entry opinion, we instead model entry comment extremism, which is a binary value (either 1 or 0). This is ultimately much more faithful to the data we have available, although it is conceptually limiting. It should be noted that part of our future work entails modeling comment generation, which will allow new agents to probabilistically generate new comments after entry. This will allow new agents to develop more middling opinion values over time, in a manner that is faithful to our limited observations.

Finally, we must consider the functional forms of our regressions. A key difference between entry and exit is that we can observe the agents who may potentially leave, because they are currently in our dataset, but we cannot observe the agents who can potentially enter. This allows us to consider the probability of exit for each agent, as a function of both agent-level and global information, in a way that is not possible with entry. Accordingly, we use a logistic regression to determine agent exit, with the dependent variable representing that probability that agent A_i leaves on any particular day. Entry, on the other hand, is represented as two global variables, each counting the daily number of people who enter and post an extreme or non-extreme comment respectively. The same global explanatory variables are used to predict both variables, via two standard linear regressions. In both cases, when trying to make a prediction for time t in the overall model, we would only use data from times before t. For the sake of providing clear and explicit examples, in the following subsections, we will accompany descriptions of the data and

model with descriptive statistics and results taken from regressions on the whole dataset.

5.5.1 Description of Data

Both entry and exit use many of the same dependant variables. However, the logistic model has one observation per agent/day pair, whereas the linear regressions, being global, each have only one observation per day. Thus, while the descriptions of the variables are the same, the descriptive statistics vary with the regression. Accordingly, we provide descriptive statistics for the two regressions separately, in Table 5.1 and Table 5.2, while elaborating on their definitions below.

First, for both regressions, we only look at days from $d = 84$ and beyond. This is because of the aforementioned behavior seen in Figure 5.4, where initial behavior is erratic.

For the two entry regressions, the dependent variables are **Number of New Entries - Extreme** and **Number of New Entries - Non-Extreme**. These are as they sound, counting the number of users who posted for the first time on day d, while grouping by whether their comment was extreme or not.

The explanatory variables for these two regressions are **Number of Active Users**, **Avg. Comment Ex.**, **Subreddit Inactive Dummies** (1 through 6), **Weekly Seasonality Variables** (Tuesday through Sunday), d, and d^2. d and d^2 are certainly the simplest, representing number of days and its square. The six subreddit inactive dummy variables are indicator variables for days when `r/Incels` became noticeably less active, due to periods where the subreddit either went private or had high levels of self-policing due to scrutiny, with the specific days listed on Table 5.1 and Table 5.2. The six weekly seasonality variables capture fluctuations in activity

throughout the week. In particular, they account for possible differential behavior on weekends versus weekdays.

Avg. Comment Ex. is slightly more nuanced, as it is *not* an average over the 28-day time-blocks we predict over, but rather a 3-day lagged average. This is simply due to testing, where we found that using the 3-day average provided a more powerful model. On any given day, this variable is the average extremism of comments from the last three days.

Finally, **Number of Active Users** intuitively should represent how many users are on the subreddit. However, because comment generation is yet to be implemented, we cannot forecast using 'number of comments' as a variable beyond even just one day. Instead, we consider a user to be active at day d so long as d is between their last and first comments. As we predict future entry and exit with these regressions at each step, we can iteratively forecast entry and exit one day forward at a time.

For the exit regression, the dependent variable is the binary indicator **Agent Left**. Specifically, if agent i is *active* on day d, with active meaning between the agent's first and last comment days, then this variable is 1 if the agent left on this day, and 0 otherwise.

The explanatory variables for the exit regression include all the explanatory variables used in the entry regressions. Additionally, this regression also includes **Time Since Joined**, **Agent Opinion**, **Agent Opinion Squared**, **Average Agent Opinion**, and a dummy variable for **First Day**. Time since joined is simply the number of days this agent has been active. Agent opinion and its square are similarly intuitive, being the agent's extremism score and its square on any given day. Each of these scores originates from the 28-day blocks, and so they only change once every 28 days. Average agent extremism is the average over all of these extremism values across all active agents. Finally, the dummy variable for first day is 1 exactly if it is one day after agent

Variable	Description	Mean	Standard deviation
New Entries - Non-Ex.	Daily number of new entrants to the subreddit who first post a non-extreme comment	64.4	50.0
New Entries - Extreme	Daily number of new entrants to the subreddit who first post an extreme comment	4.03	3.60
Active Users	Number of active users by day	1999	701
Avg. Comment Ex.	Average extremism of comments, calculated in 28-day blocks	0.0786	0.0169
Inactivity Dummies 1-6	Dummy variables for Subreddit inactivity. Covers days 102-106, 118-120, 344-348, 410-412, 432-447, and 493-498 respectively	N/A	N/A
Seasonality, Tu-S	Dummy variables to account for seasonality. Each variable covers one day of the week, from Tuesday to Sunday.	N/A	N/A
d	Time variable (day)	N/A	N/A
d^2	Time variable squared	N/A	N/A

[1] $n = 451$

Table 5.1: Descriptive Statistics for Entry Regression.

A_i's first comment. This is to help account for 1-comment users by increasing the probability of leaving after posting just 1 comment. In general, having more information for exit allows us to use more variables, with greater significance. Descriptive statistics for the entire dataset are provided in the Tables 5.1 and 5.2, with inactivity and seasonality dummy variables grouped together.

5.5.2 Regression Results

We perform the three regressions on the full dataset to obtain the regression outputs shown in tables 5.3, 5.4, 5.5. For all three regressions, we use Newey-West estimators with a lag of $\frac{3}{4} \cdot 451^{1/3} \approx 6$, to account for autocorrelation and heteroskedasticity [30].

We note that our exit logistic regression obtained a McFadden Pseudo R-Squared of 0.1867, whereas for the entry regressions we obtained Adj. R-squared values of 0.482 and 0.359 for non-extreme and extreme initial comments respectively. Moreover, nearly every variable used was significant at the 5% significance value, even under robust standard errors. The only exceptions were the seasonality variables, average comment extremism for both extreme and non-extreme entries, and active users for extreme entries. The seasonality variables were jointly significant for exit, and were preserved in the entry regression to avoid artificially induced seasonal trends in the overall prediction of population change. Average comment extremism was highly insignificant for extreme entrants, and somewhat insignificant for non-extreme entrants. Similarly, active users had a significance level of 0.083 for extreme entries, and 0.007 for non-extreme entries, which we felt still warranted inclusion in the model. Every non-seasonality variable is significant at $p < 0.001$ for our exit regression, which is explained in large part by greater availability of data. Overall, these models explain a significant percentage of entry and exit. Considering that agent entry and exit are complex events that depend in large part on unobservable real-world events, we consider these results to be very promising.

We also acknowledge that there are improvements that can be made to the model. Currently, average comment extremism is calculated over all comments. This may explain its lack of significance, as Reddit is structured so that only comments with the most upvotes are shown

Variable	Description	Mean	Standard deviation
Agent Left	Indicator variable for an agent leaving on day d	0.0455	0.208
Time Since Joined	Number of days since first comment	74.3	71.2
Agent Opinion	Opinion of agent, calculated in 28-day blocks	0.0510	0.137
Agent Opinion Squared	Opinion of agent sqaured	0.0213	0.112
Active Users	Number of active users by day	2234	473
Avg. Agent Opinion	Average Opinion over Active Agents	0.0106	0.0018
Avg. Comment Ex.	Average extremism of comments, calculated as a 3-day lagged average	0.0772	0.0205
Inactivity Dummies 1-6	Dummy variables for Subreddit inactivity. Covers days 102-106, 118-120, 344-348, 410-412, 432-447, and 493-498 respectively	N/A	N/A
Seasonality Dummies, Tu - S	Dummy variables to account for seasonality. Each variable covers one day of the week, from Tuesday to Sunday.	N/A	N/A
First Day	Dummy variable for a user's first day after joining	0.019	0.135
d	Time variable (day)	N/A	N/A
d^2	Time variable squared	N/A	N/A

[1] $n = 520900$

Table 5.2: Descriptive Statistics for Exit Regression

to a significant percent of users. Weighting each comment by visibility would account for this, and possibly produce more significant results. However, making even short-term predictions with such a model would also necessitate the creation of a robust model of comment generation, beyond accounting for just entry comment extremism, and therefore would have to be the subject of future research.

Active users may suffer from a different problem. Because the change in the number of active users is just the sum of entry and exit, using it in each model introduces auto-regressive effects. Because this variable is significant for two out of the three regressions, this indicates the presence of some amount of auto-correlation, which may be better addressed through an explicit auto-regressive model. Certainly, this would at the very least be a good topic for future investigation.

Finally, the lack of comment generation significantly hindered the development of this model. As an explanatory variable, 'number of comments' could potentially be an excellent measure of subreddit activity, and accounting for visibility within comment generation would allow for a better 'average comment extremism' metric. Moreover, it would allow us to better predict the future behavior of the agents that enter, and model change in subreddit extremism due to entry. As both of these qualities are clearly important, future work should focus largely on developing and incorporating comment generation into our regressions and overall model, as discussed in Section 5.7.

Dep. Variable:	Left_Subreddit	**No. Observations:**	520900
Model:	Logit	**Df Residuals:**	520878
Method:	MLE	**Df Model:**	21
Pseudo R-squ.:	0.1867	**Log-Likelihood:**	-82896.
converged:	True	**LL-Null:**	-1.0192e+05
Covariance Type:	HAC	**LLR p-value:**	0.000

	coef	**std err**	**z**	**P$>$ |z|**	**[0.025**	**0.975]**
Const	0.8827	0.115	7.674	0.000	0.657	1.108
Time Since Joined	-0.0272	0.001	-48.175	0.000	-0.028	-0.026
Agent Opinion	-4.3005	0.179	-24.050	0.000	-4.651	-3.950
Agent Opinion Squared	4.3678	0.199	21.931	0.000	3.977	4.758
Active Users	-0.0004	5e-05	-8.479	0.000	-0.001	-0.000
Avg. Agent Opinion	-104.5216	6.399	-16.333	0.000	-117.064	-91.979
Avg. Comment Ex.	-5.0400	0.853	-5.906	0.000	-6.713	-3.367
Inactivity Ind. 1	-5.0629	1.002	-5.052	0.000	-7.027	-3.099
Inactivity Ind. 2	-2.8473	0.502	-5.672	0.000	-3.831	-1.863
Inactivity Ind. 3	-3.3710	0.381	-8.854	0.000	-4.117	-2.625
Inactivity Ind. 4	-35.2786	0.061	-581.001	0.000	-35.398	-35.160
Inactivity Ind. 5	-4.4044	0.242	-18.185	0.000	-4.879	-3.930
Inactivity Ind. 6	-1.6600	0.126	-13.125	0.000	-1.908	-1.412
Tuesday (Seasonality)	0.0470	0.024	1.975	0.048	0.000	0.094
Wednesday (Seasonality)	0.0093	0.024	0.387	0.698	-0.038	0.057
Thursday (Seasonality)	-0.0502	0.024	-2.063	0.039	-0.098	-0.002
Friday (Seasonality)	0.0350	0.024	1.462	0.144	-0.012	0.082
Saturday (Seasonality)	-0.0580	0.024	-2.376	0.018	-0.106	-0.010
Sunday (Seasonality)	-0.0101	0.024	-0.417	0.677	-0.058	0.037
First Day Ind.	0.3294	0.030	10.831	0.000	0.270	0.389
d	-0.0087	0.001	-10.119	0.000	-0.010	-0.007
d^2	2.313e-05	1.15e-06	20.159	0.000	2.09e-05	2.54e-05

Table 5.3: Summary of Results for Exit Regression

Dep. Variable:	Daily Non-Extreme Entries	**R-squared:**	0.500
Model:	OLS	**Adj. R-squared:**	0.482
Method:	Least Squares	**F-statistic:**	55.71
No. Observations:	451	**Prob (F-statistic):**	4.20e-91
Df Residuals:	434	**Log-Likelihood:**	-2247.1
Df Model:	16	**AIC:**	4528.
Covariance Type:	HAC	**BIC:**	4598.

	coef	**std err**	**z**	**P$>$ |z|**	**[0.025**	**0.975]**
Const.	140.3681	29.571	4.747	0.000	82.409	198.327
Active Users	0.0552	0.021	2.681	0.007	0.015	0.096
Avg. Comment Ex.	-282.3241	171.748	-1.644	0.100	-618.944	54.296
Inactivity Ind. 1	-52.2317	9.326	-5.601	0.000	-70.510	-33.953
Inactivity Ind. 2	-35.8389	5.875	-6.100	0.000	-47.353	-24.325
Inactivity Ind. 3	-51.4480	6.755	-7.616	0.000	-64.687	-38.209
Inactivity Ind. 4	-82.2971	10.714	-7.681	0.000	-103.297	-61.298
Inactivity Ind. 5	-97.8971	9.917	-9.872	0.000	-117.334	-78.460
Inactivity Ind. 6	-112.4495	10.011	-11.233	0.000	-132.071	-92.828
Tuesday (Seasonality)	-0.9094	4.160	-0.219	0.827	-9.062	7.244
Wednesday (Seasonality)	-0.1477	5.416	-0.027	0.978	-10.763	10.467
Thursday (Seasonality)	-4.2755	5.103	-0.838	0.402	-14.277	5.726
Friday (Seasonality)	0.3287	4.790	0.069	0.945	-9.060	9.717
Saturday (Seasonality)	-6.0043	4.751	-1.264	0.206	-15.315	3.307
Sunday (Seasonality)	2.6575	4.240	0.627	0.531	-5.654	10.969
d	-1.1998	0.367	-3.272	0.001	-1.919	-0.481
d^2	0.0019	0.000	4.211	0.000	0.001	0.003

Table 5.4: Summary of Results for Entry Regression with Non-Extreme Initial Comment

5.5.3 Standardized Effect Sizes and Extremism Metric

Another important purpose of these regressions is to serve as a justification for our metric of extremism. In particular, by computing the standardized effect sizes for each non-indicator variable, we find that explanatory variables defined via our extremism metric have a practically significant effect. This is most evident in exit, where average agent opinion in particular explains approximately 10% of total effect (see Table 5.6), and is statistically significant to a high degree.

Dep. Variable:	Daily Extreme Entries	**R-squared:**	0.382
Model:	OLS	**Adj. R-squared:**	0.359
Method:	Least Squares	**F-statistic:**	61.36
No. Observations:	451	**Prob (F-statistic):**	4.22e-97
Df Residuals:	434	**Log-Likelihood:**	-1108.5
Df Model:	16	**AIC:**	2251.
Covariance Type:	HAC	**BIC:**	2321.

	coef	**std err**	**z**	**P$>$ |z|**	**[0.025**	**0.975]**
Const.	6.9980	1.747	4.005	0.000	3.574	10.422
Active Users	0.0023	0.001	1.734	0.083	-0.000	0.005
Avg. Comment Ex.	-2.3055	7.544	-0.306	0.760	-17.092	12.481
Inactivity Ind. 1	-3.0172	0.566	-5.332	0.000	-4.126	-1.908
Inactivity Ind. 2	-2.8617	0.495	-5.786	0.000	-3.831	-1.892
Inactivity Ind. 3	-3.3663	0.463	-7.276	0.000	-4.273	-2.460
Inactivity Ind. 4	-4.6563	0.580	-8.031	0.000	-5.793	-3.520
Inactivity Ind. 5	-5.8455	0.490	-11.918	0.000	-6.807	-4.884
Inactivity Ind. 6	-5.9137	0.546	-10.824	0.000	-6.984	-4.843
Tuesday (Seasonality)	0.5485	0.509	1.078	0.281	-0.448	1.545
Wednesday (Seasonality)	0.5979	0.511	1.170	0.242	-0.404	1.599
Thursday (Seasonality)	-0.0933	0.419	-0.223	0.824	-0.914	0.727
Friday (Seasonality)	0.2254	0.426	0.529	0.597	-0.610	1.061
Saturday (Seasonality)	-0.3876	0.390	-0.993	0.321	-1.153	0.377
Sunday (Seasonality)	0.1514	0.433	0.350	0.727	-0.697	1.000
d	-0.0614	0.024	-2.524	0.012	-0.109	-0.014
d^2	0.0001	2.97e-05	3.546	0.000	4.72e-05	0.000

Table 5.5: Summary of Results for Entry Regression with Extreme Initial Comment

Variable	Standardized Effect Size
Const.	0.8827
Time Since Joined	2.021
Agent Opinion	0.2193
Agent Opinion Squared	0.0928
Active Users	0.8936
Avg. Agent Opinion	1.107
Avg. Comment Ex.	0.3891
d	2.747
d^2	2.513

Table 5.6: Standardized Effect Sizes for Non-Indicator Variables in Exit

At the same time, our tests of inter-rater reliability as discussed in Section 3.9 also demonstrate that our metric is well-defined, and can be practically applied to real-world comments. This leads us to believe that our metric captures a significant and consistent real-world phenomenon. Additional research should be done to ensure that it accurately reflects existing notions within psychology on what characterizes extreme language.

5.6 Aggregation and Results

With all the structure of the model in place, we now evaluate its effectiveness. To do this, we train our model on our data from up to a block t and determine if it is predictive of the next block $t + 1$. Training the model involves regressing for entry and exit and performing gradient descent, as shown in Figure 5.5.

As our model is agent-based, our interpretation of results is as follows. Using data available from agents in the training period, we generate a prediction of the change in each individual agent's extremism. The results are pooled into an unlabeled set of the predicted opinion of agents in the subreddit. This pool is then altered in accordance with the predictions generated by our entry and exit regressions, removing elements and introducing new ones with probabilities

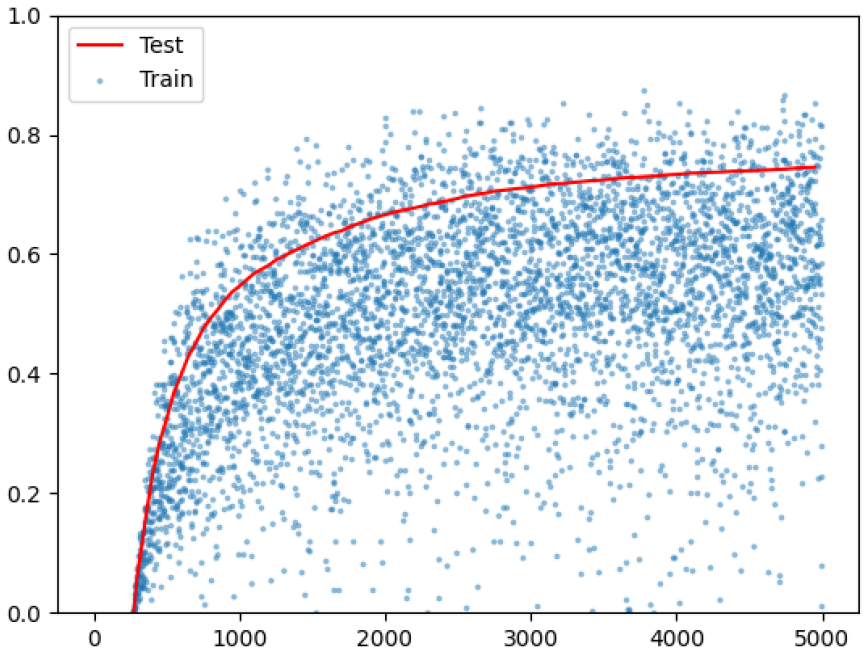

Figure 5.5: Gradient descent training curve to obtain ϕ. The plot shows r^2 of the test and training sets at each iteration. Mini-batch gradient descent was used, leading to wide variance in train set accuracy.

dictated by our regression models.

Unfortunately, the aggregated model demonstrates much less success than its component parts. This is due to the confluence of a number of factors affecting practical implementation. The foremost of these is a direct consequence of the short time scale on which our regressions must be conducted. In order to produce any long-term prediction, the entry and exit models must be applied to their own output. This, combined with the fact that the subreddit naturally grows over time, leads to extremely unstable prediction on all but the shortest scales, as error compounds and trends exponentially upwards. This is a particular challenge for our agent-based model of opinion change, which must operate on longer time scales in order to track users' posting habits accurately. Quantifying the degree to which our aggregate model fails is also difficult, as the computational load of simulating agents increases with the population's exponential growth. As such, we rely on our results taken from each portion of the model separately. In Section 5.7, we also discuss our plans to remedy these issues and refine our model to avoid these flaws.

The merits of our regressions when examined as separate entities are discussed at length in Section 5.5.2. Also of interest, however, is our agent-based model and the parameters it obtains. Through gradient descent, we obtain an interaction function ϕ which models how agents interact with the extremist content of comments based on their opinion, shown in Figure 5.6. Before drawing conclusions, we note that the gradient descent which produced this phi obtained an r^2 loss of only 0.3348, and so observations made are not rigorous.

We observe that extreme comments have stronger effects on agents with higher opinion metrics, and weaker effects on agents with lower opinion metrics. In particular, one notable trend is the negative influence of non-extreme comments on high opinion metric users, suggesting that said users are pushed away from these opinions. These observations are relatively in line with

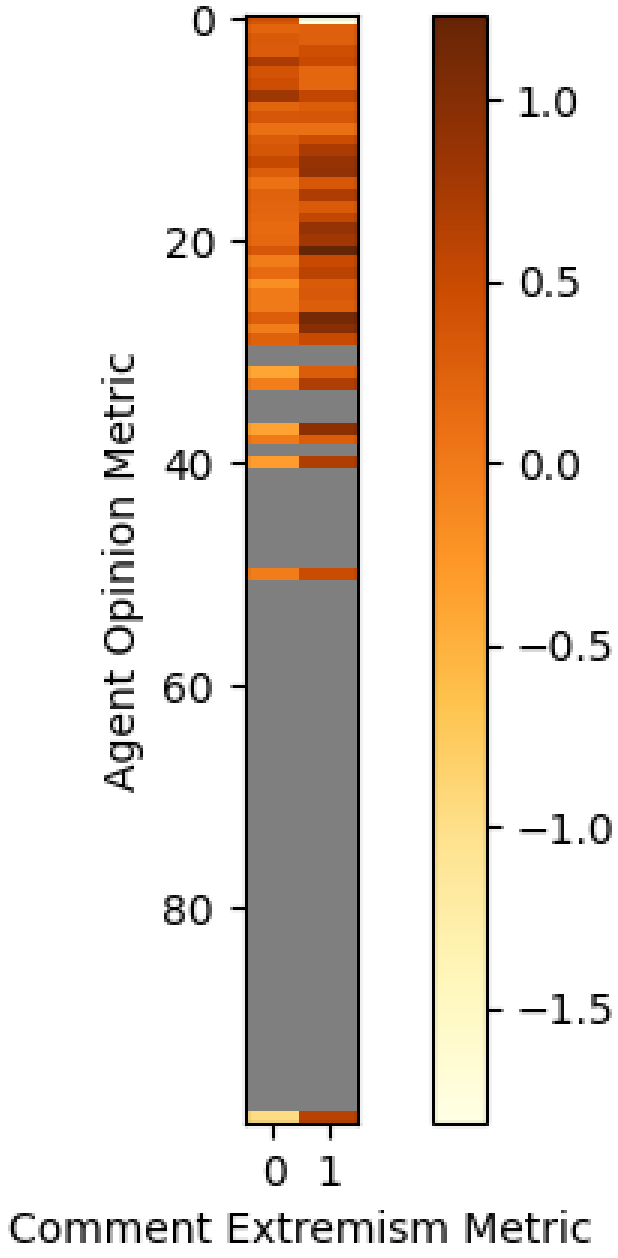

Figure 5.6: Values of phi after training using gradient descent. Values shown in grey are intentionally removed due to small sample sizes of agents with the corresponding opinion values.

intuition, although it is important to note that the bounded nature of the opinion metric scale somewhat biases towards this outcome.

5.7 Future Work

The primary area for future work on this lies in stabilizing our aggregate model by tempering its nonlinear trends of growth. We expect this to be a promising area of research, as our entry and exit regressions demonstrate significant predictive power. Once this is accomplished, our next area of work would be to attempt to construct a model of comment generation. This would allow the model to create predictions of the future without any knowledge of the upcoming comments. Furthermore, we believe this to be a very attainable goal considering the degree to which our agent opinions are predictive of the extremism of future posts.

In addition, we have plans to examine two parameters in particular: visibility for comments and engagement for agents. For both parameters, we provided adequate approximations which served our purposes, but we feel that these both leave room for experimentation and improvement. The functional forms which could be used for either parameter are numerous, and merit a deep investigation of their own.

From a broader perspective, we would like to explore more modeling and data-analytic approaches to this particular problem. For example, the consensus model upon which our model is based may be a valid approach if one examines the agents in a vacuum without the presence of comments. In terms of data, our primary area for potential improvement lies in using our BERT model to classify comment extremism with more granularity, which we anticipate will significantly increase the sophistication of our model. Furthermore, we would like to see if our

results generalize to other similar or dissimilar subreddits, including the unofficial successor of `r/Incels`, `r/Braincels`. We also would like to explore incorporating posts into our dataset, a task we previously avoided due to the multimedia nature of the data and thus the difficulty of processing posts in an unbiased and usable manner.

Chapter 6: Diversity, Equity, and Inclusion

6.1 Selection of `r/Incels`

Any claim that a particular group is extremist should be treated seriously. Social media platforms and law enforcement agencies operate harshly against the threat of extremist groups. We believe that the incel community deserves the careful scrutiny given to extremist groups. Self-described incels are responsible for several deadly domestic terrorist attacks in North America since Elliot Rodger's attack at the University of California, Santa Barbara in 2014. Alek Minassian posted online of his allegiance to the "Incel Rebellion" and Elliot Rodger before his own deadly attack in Toronto in 2018 [3]. Not every person who subscribes to the incel ideology supports violence as a means of political action, but as we have observed from real world attacks and read personally, some do.

6.2 Defining and Measuring Extremism

Ideology cannot be strictly defined, and is perceived differently by every member of an ideological community. The criteria which our team chose to define extremism are inherently imprecise. Despite this limitation, the features identified in our weighted lexicon approach and rating guide are informed by peer-reviewed descriptions of extremist narratives.

Our individual ratings of comments, while again inherently imprecise, are relatively consistent within the team as demonstrated in Chapter 3 and became more consistent with further iterations.

Taken together, this implies that our subjective measurements of the extremist content in comments are internally consistent; we are measuring an observable phenomenon with precision. Further, this observable phenomenon is based on previously published criteria of extremist narratives.

6.3 Rating Guide

Care needs to be taken in distinguishing between the extremist ideas present in the incel community from many other critiques of society. This is especially evident in the problem with the world statements category, which could possibly conflate the misogyny expressed by incels with the many historical and ongoing efforts to improve women's position in society, for example. That is why we have focused our research on a specific ideology which is intolerant, hateful, and violent. While it may potentially be modified by researchers of other extremist ideologies, our rating guide is designed to prioritize this particular set of extreme ideas and particular features that are present in the rhetoric.

6.4 Machine Learning Methods

The implementation of our machine learning methods (logistic regression and BERT) improved our classifier's accuracy at the cost of transparency. However, we still worked to ensure that our automation did not misinterpret healthy and benign discussions of faith, race, gender, or sexuality. In cases of a user revealing one such aspect of his/her/their identity, it is essential that the classifier not behave differently than if the user hadn't. For example, many of the comments manually rated by our team members purport the idea that the comment author is not able to attract women because he is non-white. Now, the futility in this sentiment (i.e. problem with the world" ideation) may call for a non-zero extremism score. However, BERT should not report a dramatically different rating for such a comment than for the very similar (and also quite common) idea that the comment author is unable to attract women because he is short.

6.5 Interpreting the Model

The mathematical model described herein is agent-based, and as such ascribes to each user an "Opinion" function which gauges the rate at which they produce comments classified as extreme by our automated extremism detection system. The model uses these functions and produces predictions about the future value this opinion function will have and thus the types of comments the user will generate. It is critical to note, however, that these predictions are not intended or expected to be accurate for individual users. Due to the high degree of randomness and unobservable change in each user's opinions outside of the context of `r/Incels`, it is not possible to meaningfully distinguish between users who will become more extreme and users

who will become less extreme. We make no claims about the reliability of our projections for individual users and condemn any attempt to use them to draw conclusions in such a context. Although this model is agent-based, its predictive power is only in considering the composition of the subreddit as a whole.

Bibliography

[1] Imran Awan. Cyber-Extremism: Isis and the Power of Social Media. *Society*, 54(2):138–149, April 2017.

[2] Robin Thompson. Radicalization and the Use of Social Media. *Journal of Strategic Security*, 4(4):167–190, 2011.

[3] Mark Alfano, J. Adam Carter, and Marc Cheong. Technological Seduction and Self-Radicalization. *Journal of the American Philosophical Association*, 4(3):298–322, 2018.

[4] U.S. Department of Homeland Security. The Internet as a Terrorist Tool for Recruitment & Radicalization of Youth. Technical report, Homeland Security Institute, April 2009.

[5] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The Pushshift Reddit Dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14:830–839, May 2020.

[6] Reddit about page: https://www.redditinc.com/.

[7] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. Regulating behavior in online communities. *Building successful online communities: Evidence-based social design*, pages 125–178, 2012. Publisher: MIT Press, Cambridge, MA, USA.

[8] Tarleton Gillespie. Content moderation, AI, and the question of scale. *Big Data & Society*, 7(2):2053951720943234, July 2020. Publisher: SAGE Publications Ltd.

[9] Robert Gorwa, Reuben Binns, and Christian Katzenbach. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1):2053951719897945, January 2020. Publisher: SAGE Publications Ltd.

[10] R. Sathya and Annamma Abraham. Comparison of supervised and unsupervised learning algorithms for pattern classification. *International Journal of Advanced Research in Artificial Intelligence*, 2(2), 2013.

[11] Saja Aldera, Ahmad Emam, Muhammad Al-Qurishi, Majed Alrubaian, and Abdulrahman Alothaim. Online Extremism Detection in Textual Content: A Systematic Literature Review. *IEEE Access*, 9:42384–42396, 2021. Conference Name: IEEE Access.

[12] Mohammad Hajarian and Zahra Khanbabaloo. Toward Stopping Incel Rebellion: Detecting Incels in Social Media Using Sentiment Analysis. In *2021 7th International Conference on Web Research (ICWR)*, pages 169–174, May 2021.

[13] David M. Blei and John D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning - ICML '06*, pages 113–120, Pittsburgh, Pennsylvania, 2006. ACM Press.

[14] C. X. Lin, Q. Mei, J. Han, Y. Jiang, and M. Danilevsky. The Joint Inference of Topic Diffusion and Evolution in Social Communities. In *2011 IEEE 11th International Conference on Data Mining*, pages 378–387, December 2011.

[15] Mehrdad Farajtabar, Yichen Wang, Manuel Gomez Rodriguez, Shuang Li, Hongyuan Zha, and Le Song. COEVOLVE: A Joint Point Process Model for Information Diffusion and Network Co-evolution. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1954–1962. Curran Associates, Inc., 2015.

[16] Astrid Bötticher. Towards Academic Consensus Definitions of Radicalism and Extremism. *Perspectives on Terrorism*, 11(4):73–77, 2017.

[17] Pierre-Emmanuel Jabin and Sebastien Motsch. Clustering and asymptotic behavior in opinion formation. *Journal of Differential Equations*, 257(11):4165–4187, December 2014.

[18] A. Wood, P. Tanteckchi, and D. A. Keatley. A Crime Script Analysis of Involuntary Celibate (INCEL) Mass Murderers. *Studies in Conflict & Terrorism*, 0(0):1–13, February 2022. Publisher: Routledge _eprint: https://doi.org/10.1080/1057610X.2022.2037630.

[19] R. Bennett Furlow and Jr H. L. Goodall. The War of Ideas and the Battle of Narratives: A Comparison of Extremist Storytelling Structures:. *Cultural Studies ? Critical Methodologies*, May 2011. Publisher: SAGE PublicationsSage CA: Los Angeles, CA.

[20] Jan-Willem van Prooijen and André P. M. Krouwel. Psychological Features of Extreme Political Ideologies. *Current Directions in Psychological Science*, 28(2):159–163, April 2019. Publisher: SAGE Publications Inc.

[21] J. Richard Landis and Gary G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159, March 1977.

[22] Klaus Krippendorff. Reliability in Content Analysis: Some Common Misconceptions and Recommendations. *Human Communication Research*, 30(3):411–433, 01 2006.

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May 2019. arXiv: 1810.04805.

[24] Javier Torregrosa, Gema Bello-Orgaz, Eugenio Martinez-Camara, Javier Del Ser, and David Camacho. A survey on extremism analysis using Natural Language Processing. *arXiv:2104.04069 [cs]*, April 2021. arXiv: 2104.04069.

[25] Rainer Hegselmann and Ulrich Krause. Opinion Dynamics and Bounded Confidence Models, Analysis and Simulation. *Journal of Artificial Societies and Social Simulation*, 5, July 2002.

[26] Ulrich Krause. A Discrete Nonlinear and Non-Autonomous Model of Consensus Formation. *Communications in Difference Equations*, 2000, July 2000.

[27] Sebastien Motsch and Eitan Tadmor. Heterophilious Dynamics Enhances Consensus. *SIAM Review*, 56(4):577–621, 2014. _eprint: https://doi.org/10.1137/120901866.

[28] Eric Silverman, Jakub Bijak, Jason Noble, Viet Cao, and Jason Hilton. Semi-Artificial Models of Populations: Connecting Demography with Agent-Based Modelling. In Shu-Heng Chen, Takao Terano, Ryuichi Yamamoto, and Chung-Ching Tai, editors, *Advances in Computational Social Science*, Agent-Based Social Systems, pages 177–189, Tokyo, 2014. Springer Japan.

[29] Andrew J. Yoak, John F. Reece, Stanley D. Gehrt, and Ian M. Hamilton. Optimizing free-roaming dog control programs using agent-based models. *Ecological Modelling*, 341:53–61, 2016.

[30] Whitney K. Newey and Kenneth D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987.