5-2019

Understanding prediction of low-quality comments in online science discourse

Elizabeth R. Lucas
erl7902@rit.edu

Understanding prediction of low-quality comments in online science discourse

by

Elizabeth R. Lucas

A Thesis Submitted
in
Partial Fulfillment of the
Requirements for the Degree of
Master of Science
in
Computer Science

Supervised by

Dr. Cecilia Ovesdotter Alm
College of Liberal Arts

and

Dr. Reynold Bailey
Department of Computer Science
B. Thomas Golisano College of Computational and Information Sciences
Rochester Institute of Technology
Rochester, New York

5 2019

The thesis "Understanding prediction of low-quality comments in online science discourse" by Elizabeth R. Lucas has been examined and approved by the following Examination Committee:

Dr. Cecilia Ovesdotter Alm
Associate Professor

Dr. Reynold Bailey
Professor

Dr. Joe Geigel
Professor

Abstract

Understanding prediction of low-quality comments in online science discourse

Elizabeth R. Lucas

Supervising Professors:
Dr. Cecilia Ovesdotter Alm and Dr. Reynold Bailey

Online public forum discussion in the Reddit science community is moderated to ensure a high standard of scientific content in discourse. This thesis project creates predictive models to distinguish between acceptable and unacceptable forum comments, and it seeks to interpret those models based on input from moderators. Information was collected from moderators (their demographics, moderation habits, moderation decisions and reasons for removing comments) to curate a deeper understanding of the online moderation community and predictive models for automated moderation support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The collection of online forums named Reddit is a website where user-generated content is submitted and voted on by the community. Reddit is comprised of individual forums known as subreddits, which focus on a shared interest or topic. Communities are usually monitored by moderators, i.e. volunteers responsible for removing unwelcome content. The Science subreddit, stylized as */r/science*, is heavily moderated to ensure scientific content and discourse. Posts with clickbait titles or a low impact factor are removed, and jokes or unscientific comments are not allowed. The Science subreddit's current moderation structure incorporates AutoModerator (AutoMod), which uses a list of keywords provided by that subreddit's moderation team, which are passed through a regular expression (regex) pattern matching system that either flags a comment for review by human moderators, or removes it [6]. While a regex-based system is sufficient to censor specific words or phrases, its rigidity makes it unable to capture many inappropriate comments and adapt to changing discourse. The AutoMod system checks every comment as it is posted, but many comments are never seen by human moderators.

1.1 Problem Definition

This thesis seeks to understand how standards in the moderation of scientific discourse can be quantified, and how to create a machine learning framework that can differentiate between acceptable and unacceptable comments. It explores two hypotheses:

H1 For */r/science* comments, a predictive model can distinguish an acceptable from an unacceptable comment, with particular focus on comment removal decisions as compared to AutoMod and reference data removals

H2 Data collected from moderators on their moderation characteristics and decision-making including comment rejections and reasons sheds light on human moderation and can be used to interpret model behaviors.

1.2 Motivation

This thesis is motivated by the understanding that online moderation is a manual, fatiguing, emotionally exhausting, and potentially traumatizing task, and that this moderation is essential to the health of the discussion community. Communities, therefore, can potentially benefit from automated moderation systems that reduce the amount of manual work for the moderators without allowing an increase in toxic content.

We are also motivated to quantify the standards of moderating scientific discourse, as determining the quality of a discussion is a subjective task. While Reddit has become a popular source for research on communities and online behavior, the nature of content in strictly and rigorously moderated communities like */r/science* have not been extensively researched. Deep learning networks have become more important to trace the reasoning behind a model's decision. This is particularly important for moderation, a socially acceptable form of censorship.

1.3 Contributions

This thesis aims to:

1. Collect a new, expertly annotated text-based dataset of moderated scientific discourse with a newly created annotation web application.
2. Analyze the collected dataset to quantify standards and annotation practices for online moderation of science discourse.

3. Create models using modern classification techniques, including standard methods as well as attention-based deep neural networks.

4. Analyze and interpret the efficacy of model architectures, combining evidence from model performance and responses from human moderators.

Chapter 2

Related Work

2.1 Scientific Discourse

Scientific English is often considered to be a specific register, or variety, of English. This register was traced through multiple historical corpora, including the documents of the English Royal Society, in which many scientists of the time favored a straightforward style of rhetoric [28]. Halliday, when discussing the specific Scientific English found in physical science literature, refers to the birth of Scientific English as Chaucer's *Treatise on the Astrolabe*, which was written in the fourteenth century [21].

Studies of the features of scientific discourse and scientific writing show that these registers have a specific structure. Within research articles, patterns have been found with respect to context frames and the establishment of marked themes [20]. Banks' work selects a number of articles from various time periods, and discusses the general increase in grammatical metaphor by way of nominal processes in scientific writing over the past 250 years [5]. These trends and structures establish scientific discourse as its own unique register within English, which may affect the way science is discussed online.

2.2 Moderating Online Communities

Online moderation, which includes the removal of comments, posts, users, and even entire communities, is an overwhelmingly manual task. Moderation on social media can be performed either by volunteers or paid workers. Moderation, in the context of Reddit, can be viewed as a civic labor in which individuals collaborate to maintain a baseline of quality and enforce guidelines laid out by the community. Moderators can view themselves and behave in a multitude of ways. Some may view themselves as *dictators*, selectively enforcing policies and rules by their own judgment, or as *janitors*, cleaning up the toxic content that other users bring to their community [25].

Moderators are also typically the first line of defense on social media for uncensored content, and can suffer traumatic effects from constant exposure to uncensored media. Employees at multiple social media companies have discussed the effects to their psyche such as shock, trauma, and desensitization to gore, sexually explicit images, and even illicit material such as child pornography [38]. Constant, repetitive exposure to conspiracy theories and fake news can also permanently shift the moderator's views of the world. While mental health resources for those that develop Post Traumatic Stress Disorder (PTSD) or Secondary Traumatic Stress may be available while working as a moderator, these resources are often not available should symptoms manifest after leaving the company [29].

Moderation and the removal (also known as banning) of communities can have positive effects on the platform, including a potential downturn in hate-speech, toxicity, and other undesirable behaviors within a social media platform. Chandrasekhara et al. (2017) found that when Reddit banned toxic subreddits Fat People Hate (*/r/fatpeoplehate*) and Coon Town (*/r/coontown*) in 2015, there was a measurable overall decrease in hate speech across the entire platform [11]. Preemptive moderator actions, such as reminding users of established rules on */r/science* has been found to increase participation, and reduce the overall amount of comments that are removed by moderators [26].

2.3 Applying Machine Learning to Linguistic Meaning

Capturing linguistic meaning within computational systems has long been an area of interest; modern research in machine translation, for example, dates back to the 1950s [37]. Topics within computational linguistics range across the formal structures of language, including parsing of syntax or semantics, and impact all linguistic levels from phonemes to full bodies of text. Word vectorization and word embeddings have been used in recent years to understand how words within a language relate to each other [27, 31], and common models for natural language processing apply techniques such as word vectorization and bag-of-words, in which texts are represented as a set of the words contained within them.

Context is an key facet within the structure of human language, and many machine translation programs have found success using models that play to this feature. Long Short-Term Memory (LSTM) models in particular are able to retain multiple layers of decisions to more accurately predict the next word based on the context of the past words in a given sequence [4, 22].

Attention-based models have been gaining traction in recent years for use in several different linguistic tasks, including text summarization and understanding. These models are able to shift their context to better focus on key words or features that have a larger impact on the overall meaning of the text or are more important for contextual understanding of a phrase. Neural attention models have been used for sentence and paragraph summarization, in which an attention-based encoder is added to a feed-forward neural network language model [33, 36], including syntax and semantics evaluation [24].

2.4 Leveraging Machine Learning and Natural Language Processing for Moderation

Researchers have explored multiple different domains to understand the characteristics of what would pass a community's moderation standard, specifically Correa and Sureka's (2014) investigation of Stack Overflow [14], and Agichtein et al.'s (2008) work with Yahoo! Answers [2]. Both are Question and Answer (Q&A) sites in which a user poses a question for the community of users to answer, and also includes a reputation system in which users can vote on content.

Correa and Sureka attempt to understand what makes a bad question on Stack Overflow, using deleted and non-deleted questions from a database. The moderators of Stack Overflow also selected from a list of reasons explaining the deletion, including *off-topic*, *not a real question*, and *subjective*. The authors found that a mix of features are important to properly classify a question, with the most distinguishing features being the history of the user's account. Without relying on previous question and answer history, they found that certain wordings could point towards the classification of the question as acceptable or unacceptable.

These features are also present for the models created for Yahoo! Answers, including the presence of punctuality and typos. Agichtein and colleagues focused more on the quality of the text within the question, including other metrics such as readability, which they determine by the average number of syllables in words and their lengths. Both systems found that their models improved when using a mixture of textual features and information about the community at large, such as user history and leveraging the reputation metrics of the platform.

The use of machine learning models to understand content quality can intuitively be extended to leverage such quality metrics for censorship of toxic content or moderation purposes. Previous work on automated moderation has often used bag-of-words approaches [10, 12]. The bag-of-communities approach iterates on the concept of bag-of-words and determines quality by the content's similarity to pre-selected communities [12]. The communities have been selected by the authors as indicative of either toxic and unmoderated communities, including 4chan boards (a collection of image-based forums known for their toxicity, vulgarity, and lax moderation) and certain hate-oriented subreddits, or as well-moderated (i.e. *model*) communities, including several subreddits. The content from these strictly moderated subreddits serves as a standard that reflects the quality the authors wish to see for their target platform. All communities have their content scraped and features extracted, but are given no other labels in terms of content quality besides the community from which they originate. The authors were able to successfully quantify a text sample's similarity to model versus problematic communities. These metrics could then be used to classify text from a separate platform as acceptable or unacceptable.

Non-Q&A platforms have also been researched, including Instagram. Instagram is a photo-based platform, where users post a photo and optionally include a caption as well as tags (known as *hashtags*) that are searchable. By focusing on tags that represent disallowed material, in this case content that is favorable towards eating disorders (ED), Chancellor, Lin and De Choudhury (2016) were able to train a model to recognize posts with this problematic content with 69% accuracy using a bag-of-words approach [10]. Compared to the baseline, this model could be used as a supplementation to current moderation methods in order to catch a wider range of content. A mixture of caption and tag unigrams were used to create features for the posts, and the researchers additionally found that some key phrases associated with pro-self-harm, another topic that is not allowed on Instagram, were prevalent in ED posts.

2.5 Summary

The body of work on related topics to this thesis both motivates and demonstrates the difficulty of the moderation task, both for human moderators and automated systems, and the large amount of data needed to create an effective model. Moreover, it asserts the need to increase knowledge of human moderation habits, especially for more experienced moderators, and to leverage these insights in developing or evaluating automated moderation systems. Cleaning up online platforms provides an invaluable service to the rest of the community, protecting users from hate speech, toxic users, and triggering content such as eating disorder and self harm content. The emotional and time cost of moderating these communities, often for little to no pay, highlights the necessity of automated systems and the importance of this work.

Chapter 3

Data Collection Experiment: Moderator Survey and Moderation Annotation Task

3.1 Experimental Design

Given that science discourse moderation is a challenging and subjective task for humans, to ensure quality data and to better understand */r/science* moderation and moderators' reasoning habits about scientific texts, two instruments were used to collect data from human moderators:

1. A Qualtrics survey to collect demographic and moderation habit data from experienced human moderators.
2. A moderation annotation task completed in a new web application implemented by the researcher to label (annotate) comments with *pass* or *remove* and the reasoning for an individual comment's removal.

This study was IRB approved. Participants were recruited from the science subreddit's moderation team. Twelve people completed the moderator survey, and ten completed some amount of annotations within the annotation task, which was a limitation of the study. The overwhelming majority of moderators had completed a formal education in their area of expertise, many with advanced degrees, and they came from a variety of backgrounds and experiences.

3.1.1 Survey about Moderation

The survey provided an opportunity to holistically assess the variation in moderation habits across participants, and in reasoning for removal. Results were assessed with analysis by moderation demographics, inter-annotator agreement metrics, and visualizations of data from open-ended responses. This includes considering demographics such as age, gender, broad geographic location, and moderation-specific questions such as years of moderating service on the science subreddit, rank within the moderator hierarchy, and degree field.

Open-ended questions on moderation habits sought to understand how moderators both look for comments and decide how those comments are removed. This information could later be used to provide insight into what constitutes a high quality or acceptable comment on */r/science*, and help to understand the architecture and prediction behaviors of the generated machine learning models.

3.1.2 Web Application for Moderation Annotation Task

A newly developed web application was used to collect labels for a random selection of reddit comments. The participants were asked to judge a given comment as acceptable (pass) or unacceptable (remove), and further prompted to provide a reason, an explanation for their decision, and to elaborate if selecting *not scientific*.

The web application was written using the Flask Python library, and calls stored procedures within a MySQL database[1]. The website was hosted on DigitalOcean and Namecheap. Screenshots of the annotation application can be found in Appendix B, showing the annotation task interface and specific questions. The database has several tables: user information on web application, comment information (comment text, who removed it and when), the user annotation responses, and the batch reference list, which divides the comment dataset into multiple batches to ensure coverage in the annotation procedure.

[1] Original template: `https://github.com/jay3dec/PythonFlaskMySQLApp_Part7`

3.1.3 Overview of Comments and Data Collection Experiment

Reddit comment data was leveraged from an existing public dataset on BigQuery, comprising billions of comments [1]. For the year of 2017 alone, the dataset contains roughly 260 GB of data, including *>1M* unique comments from */r/science*. This dataset was augmented by scraping the subreddit for comments that have been removed by the moderation team. The augmentation of this dataset with removed comments aimed to mitigate the amount of dangling references from moderator and Reddit administrator intervention. After completion of the survey and task by participants, their labeled comments, including removal reasons, from the task were stored and compared to the original moderator action from the public dataset.

The limitations of the public dataset, also hosted on Pushshift, have recently been published [19]. The risk to a given Reddit user's data being incomplete due to gaps in data collection is roughly 4%, but overall only 0.043% of all comments ever published on Reddit are missing from this dataset. Gaffney and Mathias (2018) discuss the primary concern with using such a dataset for machine learning models, in particular the dangling references from removed and deleted comments. The augmentation of this dataset in the study from scraped comments, as well as the extremely low percentage of comments missing from the dataset, suggest minimal to no risk to the validity of this study due to data collection gaps.

Ten thousand comments were randomly selected from all comments submitted to */r/science* over the past five years, based on the public dataset described above. This included both top-level comments, and those that were within a thread with potentially parent and children comments. After removing automated comments, such as spam or moderator notifications, roughly nine thousand from the original pool were used for the machine learning models, and a subset of these comments were selected for use in the annotation task, which, after annotation, became part of the test data.

In order for the data subset for the annotation task to reach a ratio of 75% Removed comments and 25% Not Removed comments, several thousand comments were excluded. After balancing and removing automatically generated comments, the total sample size for annotation in the study was 3,394. These were split into batches of 200, presented to participants in sets of 100, maintaining the ratio of removed to not removed comments. Participants were asked to complete a minimum of one batch.

Two individuals who responded were excluded from analysis due to lack of qualifications or revoked consent. Out of the eligible responses, twelve participants completed the survey, and of those twelve, ten further contributed to over a thousand annotations combined in the task discussed in section 3.3.

3.2 Moderation Survey and Resulting Demographics

For the exact wording and presentation of the survey questions, please consult Appendix A. Tables 3.1, 3.2, 3.3, and 3.4 describe the distribution of both the general and */r/science*-specific demographic data collected from responding moderators.

As noted in Table 3.1, participants were skewed towards a younger audience, with the majority within the 25-34 age bracket. This could have an effect on the highest attained education of the respondents, as many could still be working towards their doctorate degree. The gender split of the participants was fairly even, with none self-identifying as a non-binary gender, and all were located within the United States. All but one self-identified as white.

The */r/science*-specific demographic questions aimed to understand the moderation habits as well as the distribution of experience within the participants. As seen in Table 3.2, an overwhelming amount of moderators were of the lieutenant rank, which entails expanded duties beyond the limited scope of comment moderators; however, they must report to the full moderators, who have more responsibility. Both Lieutenant and Full moderators are responsible for removing posts that violate the rules, and responding to messages that redditors send to the moderator team. This shows a significant prominence in respondents toward higher ranking moderators. Although comment moderators make up the vast majority of the volunteers, the majority of moderator actions (such as removing a comment or banning a user) performed by humans are authored by lieutenant and full moderators, as

Variable	*N = 12*
Age	
18-24	1
25-34	7
35-44	3
45-54	1
Gender	
Female	5
Male	7
Ethnicity	
Asian	1
White	11
Highest Attained Education	
HS or equivalent	1
Bachelor's	3
Master's	3
Doctorate	5

Table 3.1: Demographics of survey respondents. Most respondents were 25-34 years old, white, and had completed higher education. The sample was roughly gender balanced.

there are no enforced quotas for comment moderators to perform a certain number of moderation actions in a given time. The participants of this study therefore are more likely to spend significantly more time moderating */r/science* and more experienced than the regular comment moderator.

Table 3.3 reveals results on reported moderation style methods. In this context, *passive browsing* refers to the act of browsing the subreddit comment threads, as they would any other part of Reddit, and removing rule-breaking comments as they are read. *AoS pings* refers to notifications that are sent the private Slack group chat, which contains a subset of moderators that have opted-in to that communication service. If a moderator reports a

Moderator Demographics	*N = 12*
Rank	
Comment	1
Lieutenant	8
Full	3
Years Spent Moderating	
6mo - <1 yr	1
1-3 years	7
4-6 years	3
7+ years	1
Hours/Week Spent Moderating	
<1 hr	1
1-3 hrs	5
3-5 hrs	4
6-10 hrs	2

Table 3.2: Moderator rank, experience, and time spent moderating. Most participants were experienced and worked several hours per week.

thread on Reddit as needing additional moderation, a notification is sent to all moderators on Slack with a link to the offending thread. If the respondent chose *other*, they were asked to explain what other moderating styles they use. One user specified that they paid extra attention to submissions that they had personally submitted, while another said that they scan new submissions to the subreddit to mitigate the influence of bad comments before they dominate the conversation. Passive browsing, pings from the AoS Slack group, checking popular threads (threads that generate a lot of traffic and by extension, comments), and checking threads in which the moderator has some experience were the most popular methods. The use of passive browsing as the main form of moderation relies on moderators seeing many of the posts to */r/science* in their usual Reddit feed, and taking the time to click through. This most likely contributes to the large amount of comments that are never seen

Moderation Style Methods	*N = 12*
Passive browsing	12
Modqueue/reports	7
Slack/AoS pings	11
Popular threads	10
Area of expertise threads	10
Controversial threads	6
Threads likely to have rule-breaking comments	8
Other	3

Table 3.3: Responses to the survey question: *How do you moderate? Select all that apply.* Passive browsing was the most frequent moderation style method, closely followed by Slack/AoS pings, popular threads, and area of expertise threads.

by human moderators, the need for over 1,500 moderators for this subreddit, and motivates this study's work towards better automated moderation support.

The participants were also asked to approximate how much of their time is spent using each moderation method. In some cases, the participant did not select 0-19% if they did not use that method, and instead selected none of the options leading to less than 12 results for some styles. While 80-100% was also an option, none indicated they used any one style more than 60-79% of the time, so that category was excluded from Table 3.4. This table shows the wide variation of how moderators spend their time moderating across methods. While many moderators use multiple methods, each person has a different approach to what kind of threads or moderation tools they focus on, and this could help the moderation team to have a wider coverage of the thousands of comments posted to *r/science* at a given time. The results also indicate that while the AoS Slack group pings do not take up much time for most moderators, presumably because it is not pinged often, nearly every moderator surveyed spends at least some part of their time checking the threads from the pings, suggesting that this is an effective way to receive additional coverage from experienced moderators.

% Moderation using each style	0-19%	20-39%	40-59%	60-79%
Passive browsing	7	2	2	1
Modqueue/Reports	4	3	1	1
Slack/AoS Pings	10	0	0	1
Popular Threads	2	6	1	1
Area of Expertise Threads	5	4	0	1
Controversial Threads	3	4	2	0
Threads likely to contain rule-breaking comments	4	4	0	1
Other	3	1	0	1

Table 3.4: Responses to survey question: *Approximately what percentage of your moderation comes from each method?* Slack/AoS pings and passive browsing and popular threads were estimated to take no more than 19% or 39% of moderators' time, respectively.

In addition to listing their preferred moderation methods, open-ended responses were also collected for two questions:

- Describe your methodology for deciding whether a comment should be removed.
- What kind of comments cause you to ask for a second opinion?

While some moderators strictly follow the letter of the rules when deciding whether to remove comments:

> "If I come across a comment that seems suspect, I go through the commenting rules to see if it explicitly breaks one of them. If it does, I remove."

others dig deeper into the contributions as well as the greater context both within the thread and within Reddit as a whole:

> "...If it is borderline does this comment bring value to the subreddit? If it is borderline check the users history to gauge intent of the comment."

and another participant described their more strict definition of the *Off topic* rule:

> "...I tend to be very conservative in my assessment of what constitutes a high effort post, and so I tend to remove many comments that, while not necessarily harmful, seem to add nothing specific to the conversation about the paper."

Although user history and context was not available to moderators for this study, this variation in moderation style from a strict interpretation of the rules to a loose interpretation that may expand beyond what others would remove could account for the disagreement between moderators, and in turn could affect the performance of the ML models.

The ability of the moderators to correctly determine the intent and meaning of a comment is additionally limited both by their academic knowledge and the knowledge of current trends in internet culture. When asked what comments cause them to ask for a second opinion, one participant responded:

> "...Comments that seem like they might be bigoted, but I'm not sure because keeping track of all of the new bigoted jokes/memes is more than a full time job."

As Internet culture and slang changes so rapidly, it becomes difficult for individuals who are not immersed in it to keep track of new words and phrases as they gain popularity. Many of the */r/science* moderators have other obligations including personal, school, and work obligations, limiting their ability to keep up with these trends. This can also be reflected in the AutoMod keyword listing, as this has to be maintained and updated by human moderators. When popular movies are released, for instance, Reddit users try to post spoilers for other users. In response, several phrases are included to AutoMod, such as character names, to combat these comments. These keywords and phrases are often removed several weeks later, and contribute to the constant adaptation and maintenance of AutoMod.

3.3 Moderation Annotation Task and Resulting Analysis

The annotation task application asked the following of every Reddit comment the participants saw:

- Whether they would *pass* or *remove* the comment, and why;
- If they removed, for what reason;
- If the reason was *not scientific*, to explain why.

1,021[2] annotations were collected from the ten participants, and out of these 729 were marked *remove*. Table 3.5 describes the distribution of removal reasons. The *N/A* choice refers to the case where the annotator did not choose a reason for their removal, which indicates either that the given reasons did not fit, or an error occurred when submitting.

Removal Reasons	*N = 729*
Off topic	33 %
Joke	29 %
Anecdotal	15 %
Not Scientific	12 %
Offensive	8 %
N/A	1 %
Medical Advice	1 %

Table 3.5: Distribution of removal reasons, rounded to nearest whole percent. Around one third of provided reasons involved an Off Topic or Joke-related decision.

As providing an explanation for their decision was optional, 701 out of 1021 annotations (70%) included explanations. The following word clouds represent the most frequent unigrams present within the text boxes provided to participants to expand on annotation decisions. For each annotation, the participants were asked to explain their reasoning behind

[2] As one comment was annotated twice by the same annotator, this additional annotation was excluded from future analysis

their decision, and if they selected *not scientific*, asked to further explain why that particular comment is considered not scientific.

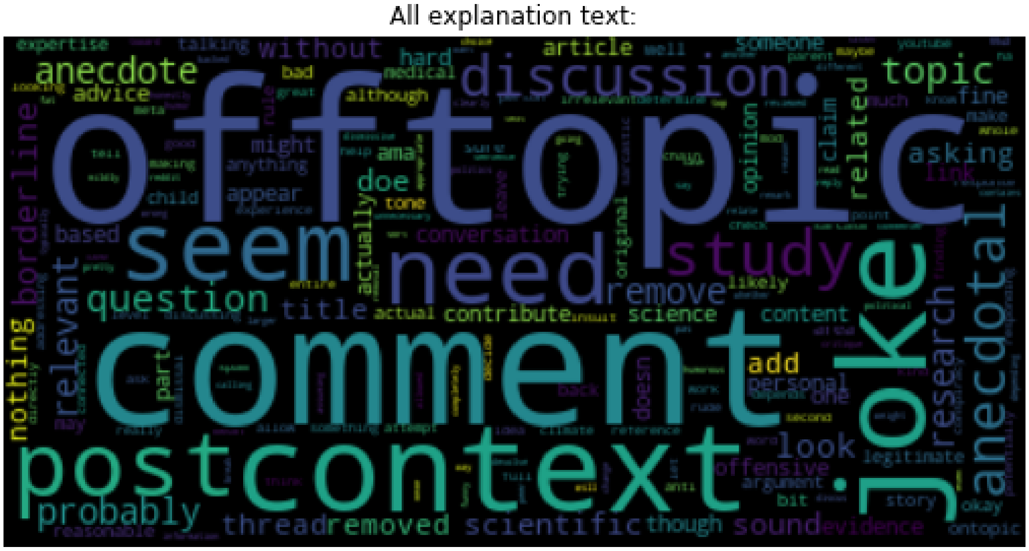


Figure 3.1: A word cloud of all text across annotation responses included within the *explanation* box. N = 701 responses. Prominent words are *joke*, *offtopic*, and *context*.

The word cloud of the explanation text in Figure 3.1 shows several removal reasons, such as *anecdote*, reflecting the *anecdotal* removal choice, as well as the high frequency of the word *joke* and *offtopic*. Participants also included within their explanations key words and phrases such as *context*, *question*, and *discussion*. The prominence of *seems*, *comment*, and *borderline* hint at the subjective nature of the moderation task.

By filtering annotation responses in which a participant disagreed with the original label of the comment, the words most frequently used changes to reflect the more difficult decisions and increasing subjectivity of the task. As shown in Figure 3.2, the word *context* appears again, referring to the context of the comment; while some comments were top-level comments, others were nested within a comment thread and are more difficult to determine without the surrounding parent and child comments. The importance of context is further highlighted by the higher frequency in Figure 3.2 in comparison to the overall explanation text in Figure 3.1; as comments become more difficult to judge, context can help the moderator to decide. The word *borderline* also rose in prominence, as comments

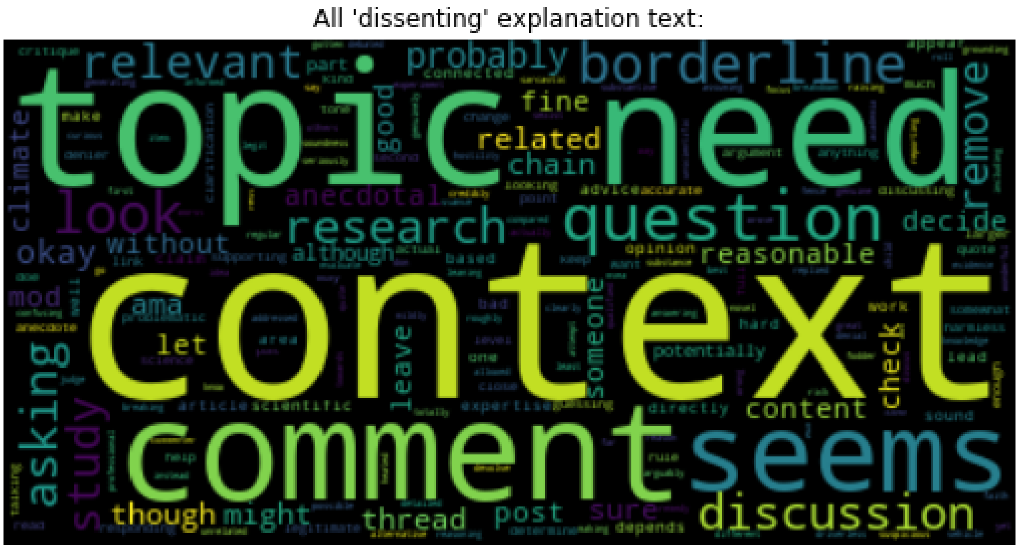


Figure 3.2: A word cloud of all text included within the *explanation* box for responses where the annotator disagreed with the original label. N = 107 responses. For these explanations, *topic*, *context*, *seems*, *borderline*, *comment*, and *question* are prominent

that are harder to judge are often referred to as borderline, i.e. on the border between acceptable and not acceptable.

One additional point of interest was understanding what makes a comment *not scientific*. The word cloud in Figure 3.3 has a much different frequency distribution than the other text responses, showing a strong skew towards *science*, *evidence*, *claim*, and *research*. This suggests comments that made claims without any evidence or sources were considered not scientific.

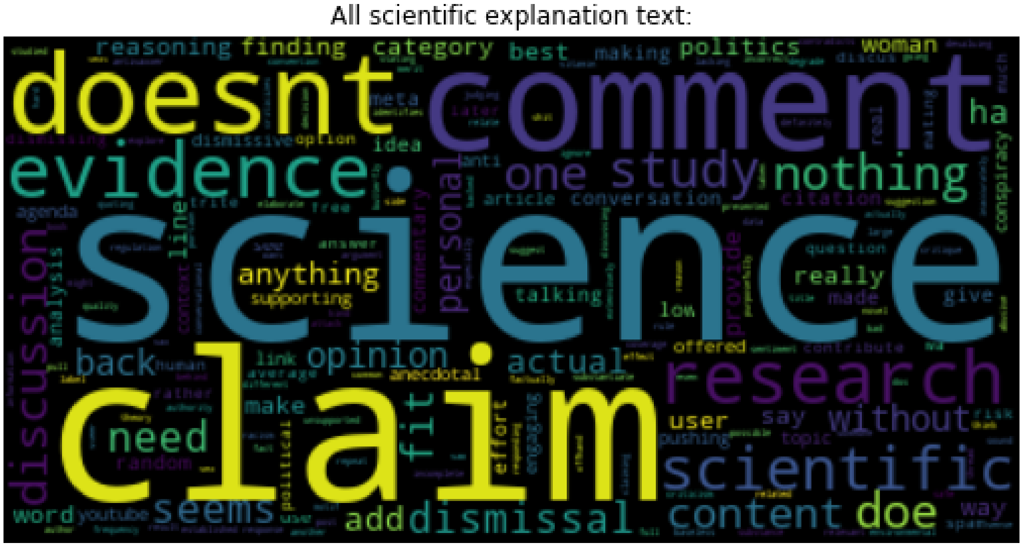


Figure 3.3: A word cloud of all text across annotation responses included within the *if not scientific, explain* box. N = 90 responses. The words *science*, *scientific*, *evidence*, *research*, *claim*, and *study* hint at attention paid by moderators to the lack of scientific rigor.

Individual explanations provided deeper insight into the specifics of how moderators decide how to remove comments. The following quotes are explanation text from annotations; many provide context for the specifics of why a comment falls under a certain removal reasons, including certain phrases that are also included in */r/science*'s AutoMod, such as */s*. These phrases often do not make sense outside of certain Internet culture context, and highlights the importance of human moderators maintaining the AutoMod keywords:

> "any comment with /s in it should be removed. /s means a comment is being sarcastic, so it's a joke."

However, there are some instances where AutoModerator removes allowed comments; this can be due to the nuance of the content itself, such as using profanity within a more acceptable context, or when a special topic such as an Ask Me Anything (AMA) thread where the rules for posting slightly change. Several comments from AMAs were included within the annotation task, and one was incorrectly labeled by AutoModerator. In one such case, an annotator selected *pass* on a comment removed by AutoMod:

> “This is for an AMA: there are different rules for allowed comments/questions in an AMA compared to regular /r/science posts.”

While not directly included in the official list of rules, some comments are considered *meta*, or talking about */r/science* instead of the particulars of the posted content. As users are able to see that comments have been removed - the text of the comment is replaced with the word *[removed]* - they often complain about how many comments have been removed. These are considered *meta* and *offtopic*, and were removed by annotators:

> “meta comment about how we remove a lot of comments”

In some cases, the annotators alluded to the practice of *nuking*, where entire chains of comments are removed due to arguments or continual off-topic remarks. The comment this annotator is referring to was not originally removed, which could indicate that a human never saw it, or that within the greater context it was not deemed inappropriate:

> “Even though it’s a nice comment, I would assume, based on the context, this entire thread should probably removed. I usually remove slap fighting.”

Additionally, as highlighted by the prevalence of the word *context* within the word clouds, this was frequently mentioned by moderators in their explanations:

> “need context, could be okay”
>
> “Borderline, need context to determine”
>
> “need context to properly decide in this case”

While some annotations where context was mentioned matched the original decision, there were also cases where the annotator chose to *pass* a comment that was originally removed by a human moderator. In the case where an annotator removed a comment that was originally not removed, we cannot determine whether a human saw the original, but if it was originally removed by a human and passed by an annotator, it suggests disagreement which could hinge on the inclusion of context.

3.4 Integrated Survey and Annotation Task Analysis

Demographic data collected from the survey was used to slice the annotations several ways by the moderator's years of experience, time spent moderating per week, and moderator rank.

	Comment Moderators	Lieutenant Moderators	Full Moderators
Moderator N	1	6	3
Annotation N	108	747	166
% Dissent	23	29	29
% Explanations	68	65	82
% Removed	69	71	72

Table 3.6: Basic statistics broken down by moderator rank. The full moderator participants provided more explanation about their reasoning.

Dissent, in the context of Table 3.6, refers to a decision made by a participant that does not match the original label for the comment. Annotation N refers to the amount of annotations each rank of moderator contributed. The full moderator participants had a much higher rate of explaining their reasoning, which may have influenced the text within the word clouds in Figures 3.1, 3.2, and 3.3. Participants across the board labeled comments as Remove at a slightly lower rate than the original label (75% of comments were originally labeled as removed, and 70% of all annotated comments were labeled removed).

Table 3.7 shows the distribution of removal reasons for individual annotated comments as broken down across moderator ranks, and the distributions are similar for the lieutenant and full moderator ranks. The second row refers to the total number of removed comments for each moderator rank. Likely because only one comment moderator was included among the participants, the distribution of removal reasons is different than the other classes, and this may additionally be caused by the comments all coming from the same batch, not benefiting from other comment batches to average out the types of comments annotated. Across all ranks, the majority of comments removed were labeled as either Off topic or

	Comment Mods. *(N = 1)*	Lieutenant Mods. *(N = 6)*	Full Mods. *(N = 3)*
Removal Reasons	N = 75	N = 534	N = 120
% Not Scientific	5	14	12
% Off topic	53	31	32
% Anecdotal	9	17	8
% Offensive	5	8	11
% Joke	27	29	34
% Medical Advice	0	1	3
% N/A	0	2	1

Table 3.7: Removal reason distributions by moderator rank in percent, with *moderators* abbreviated to *mods*. Lieutenant moderators provided the bulk of the annotations, and across all groups, Off topic and Joke were the most common removal reasons.

Joke. These accounted for roughly 60% of all comment removals for both lieutenant and full moderators, which mirrors the results shown in Table 3.5. Interestingly, when there was a large variation in the percentage of a removal reason, it occurred in only one of the three groups, such as the *comment moderator* group for the *% Off topic* reason. This could be explained either by chance in the distribution in comments that the participants saw, or by different levels of experience and time investment influencing moderation style and practices.

Results were also broken down by years of experience, shown in Table 3.8, with the majority of moderators falling into the 1-3 year category. Due to an uneven distribution of participants, the fringe categories show substantial variation in the results. However, when inspecting participants' responses, the Off topic and Joke categories have a dramatic difference in percentage as the preferred reasons for moderators of 1-3 and 4-6 years of moderation experience, respectively. This could indicate either that the two categories have substantial overlap, that moderators who joined at certain times have different definitions of what comments fit these categories, or that the motive of reasons evolved with increased experience.

	6m - 1yr *(N = 1)*	1-3 yrs *(N = 5)*	4-6 yrs *(N = 2)*	7+ yrs *(N = 1)*
Removal Reasons	N = 133	N = 426	N = 163	N = 7
% Not Scientific	11	14	10	29
% Off topic	16	42	24	57
% Anecdotal	28	13	7	0
% Offensive	11	6	9	0
% Joke	34	22	47	14
% Medical Advice	0	1	2	0
% N/A	0	2	1	0

Table 3.8: Removal reason distributions by years of experience, rounded to nearest whole percent. The moderators with 1-3 years of experience had Off topic as a more common reason, whereas respondents with 4-6 years of experience marked the Joke reason more frequently.

Finally, moderators were separated by time spent per week moderating, and the results are shown in Table 3.9. Although there were fewer participants that spend over six hours moderating per week, they provided approximately as many annotations as the other groups. While some categories are fairly consistent across all groups, such as Off topic and Medical Advice, others varied.

	1-2 hrs/week *(N = 4)*	3-5 hrs/wk *(N = 4)*	6-10 hrs/wk *(N = 2)*
Removal Reasons	N = 269	N = 210	N = 222
% Not Scientific	20	8	8
% Off topic	32	33	35
% Anecdotal	14	8	23
% Offensive	8	33	8
% Joke	23	40	26
% Medical Advice	2	1	1
% N/A	2	1	1

Table 3.9: Removal reason distributions by time spent moderating per week, rounded to nearest whole percent. The respondents reporting moderating 3-5 hours a week more often indicated the Joke removal reason.

3.5 Inter-annotator Agreement Evaluation

Due to the modest amount of participants and annotations, each comment was annotated at most twice. Out of the 1,020 annotations over 798 comments, 222 of those comments were annotated by two different participants. Using this subset of 444 annotations, the researcher examined the quality of the participants' labeling using the inter-annotator reliability metric Fleiss' kappa (κ) [18]. A generalization of the π statistic, Fleiss' κ is defined as

$$\kappa = \frac{\overline{P} - \overline{P_e}}{1 - \overline{P_e}}$$

where

$$\overline{P} = \frac{1}{N}\sum_{i=1}^{N} P_i$$

and

$$P_i = \frac{1}{n(n-1)}\sum_{j=1}^{k} n_{ij}(n_{ij}-1)$$

in which our case N represents the total number of annotations, $i = 1, 2, ..., N$ represents an individual response, and $j = 1, 2, ..., k$ represents a selection an annotator has made. $\overline{P_e}$ represents the probability that agreement between annotators happened by chance.

While we cannot determine whether a comment that was not removed was seen by a member of the moderator team and deemed acceptable or never reviewed, any comment that was removed was either removed explicitly by a moderator, or by the AutoMod system. This reference data can be used to determine both the performance of AutoMod and how often the participants of the task agree with individual assessments made by moderators. If there is high agreement within the participants and the result clashes with the original label from the dataset, we can assume that the initial judgment was made in error.

The Fleiss' κ score[3] for the subset of twice-annotated comments was 0.46, showing a modest agreement between the two annotators. As Fleiss' κ is intended for datasets with

[3] The script for determining the Fleiss' kappa score was adopted from `https://gist.github.com/ShinNoNoir/4749548`.

a larger number of annotators, Cohen's κ was also calculated using scikit-learn's built in function. Cohen's κ returned a similar result for the two annotators at 0.45.

Along with inter-annotator agreement between the participants, the analysis also measured how often the participants agree with the initial assessment of a comment being acceptable or unacceptable. This was assessed by comparing the majority response for each twice-annotated comment to the reference label in the dataset, treating them as the third annotator for the purposes of the Fleiss' κ calculation. This inclusion lowered the score to 0.33, showing lesser agreement between the two annotators and the original label than between the annotators alone.

3.6 Summary

The inclusion of an annotation task and subsequent dataset aims to better evaluate the performance of the models discussed in Chapter 4, as well as to further the understanding of the moderation of online scientific discourse and how the perception of appropriate discourse can vary between even expert annotators. As shown in this chapter, moderation styles and standards can vary greatly between people, even when accounting for differences in experience. This has been considered one of the strengths of the */r/science* moderation team, as hundreds of moderators all looking at comments with a different perspective could lead to higher quality moderation and discussion of the rationale behind decisions.

Chapter 4

Machine Learning Experiments

Linguistic features were extracted from the raw text using current natural language processing techniques. Traditional machine learning methods including Gaussian Naive Bayes (GNB), Decision Trees (DTs), and Support Vector Classifiers (SVCs) [9] were compared to deep learning models, and deep learning classifiers using attention-based models were developed for the binary prediction problem [22, 23]. The frameworks Tensorflow, Keras, and scikit-learn (sklearn) were used to develop models [17, 35]. These models were evaluated using standard metrics against the baselines of the % Removed class and the % of comments AutoModerator removed in the test set to determine their effectiveness.

To ensure a wide range of features for the models, several methods were used to vectorize the text. For various versions of the models, two different techniques were used: count vectorization, and GloVe word embeddings. Count vectorization uses the frequency of the words within the corpus to determine the number assigned to each word, and the text is converted to those numbers. The GloVe embeddings [31] are similar to word2vec representations [27], and include dense vectorization of a large number of words. For the neural network models, the GloVe corpus including 6 billion unique words was leveraged, with 100 dimensions per word vector. The machine learning pipeline was written entirely in Python, and ingested tables from the mySQL database. The data was then extracted into features using the NLTK [7] and sklearn packages [30], where it was either sent to sklearn or to Tensorflow [17] depending on the model that was being trained.

4.1 Model Architecture

The following traditional models were created using the tools in the scikit-learn library. Each model was trained using sets of unannotated comments. This was split in two parts, with a randomized 80% used for training and the remaining 20% for validation. The models were evaluated over the validation set from the original label and the annotation subsets described in Section 4.2. The vocabulary features were generated using the scikit-learn Count Vectorizer, and the word vectors were passed directly on to all three traditional models:

- Gaussian Naive Bayes (GNB)[1]

  This algorithm uses Bayes' Theorem to assume that the data has a Gaussian distribution, and parameters are estimated using Maximum Likelihood Estimation (MLE).

- Decision Tree (DT)[2]

  The DT creates a model that classifies using decision rules created from features in the training data.

- Support Vector Classifier (SVC)[3]

  By viewing the data in a multi-dimensional space, where the number of dimensions depends on the number of features used, hyperplanes can be used to divide this space. These hyperplanes become the decision boundaries, and predictions are given a label depending on how it falls.

Models were improved using scikit-learn's GridSearchCV method, which performs cross-validation and hyper parameter tuning. Each model for each experiment was individually tuned using this method. As the scikit-learn library does not allow for customization of the GNB, this approach was excluded from further adjustment using GridSearchCV.

[1] https://scikit-learn.org/stable/modules/naive_bayes.html

[2] https://scikit-learn.org/stable/modules/tree.html

[3] https://scikit-learn.org/stable/modules/svm.html

Both the DT and SVC used the scikit-learn GridSearchCV method for hyper-parameter tuning and cross-validation, and saw significant results over the out-of-the-box defaults for all of the models. The grid search used AUC and accuracy to find the best model. Originally, they all had a strong bias toward correctly classifying Not Removed (pass) comments at the expense of accuracy of the Removed class, as the original dataset is roughly split 75% Not Removed and 25% Removed, but this was lessened by balancing the classes using scikit-learn's built-in class weight balancing, which punishes the models more harshly for incorrectly labeling data from the minority class tham the majority. By weighing the Removed label heavier, while both labels were predicted, the increase in performance for the Removed label was achieved at the expense of overall accuracy given the distribution in the test set.

Two neural network models were created for comparison against the traditional models, with the first serving as a benchmark to show the difference between using a simple dictionary mapping, which creates a dictionary of the vocabulary of the test set and applies those numbers to each word within the text, and the GloVe embeddings. With a basic feed-forward neural net of few layers, there was no significant difference between the two methods. Both models performed at roughly 75% accuracy, and was achieved after 30 epochs for the dictionary mapping and 10 epochs for the GloVe embeddings. However, this value was caused by an issue in weighting the classes, and resulted in models that selected the Not Remove label for every sample.

By expanding upon the model using the GloVe embeddings, an attention metric was included [36, 24]. The architecture was based on Yang, et al.'s Hierarchical Attention Network (HAN) from 2016 [40]. This hierarchical approach encodes the sentences found within the text separately than the individual words, and trains an attention layer for both. The authors provide a diagram for the model within their paper, which reflects the architecture shown in Table 4.1.

The initial attention model has the structure as shown in Table 4.1, with only a handful of layers. However, as opposed to Yang, et al.'s description of their HAN, this network uses sentences and the full text as the two hierarchies[4]. The first four layers belong to the sentence decoder and attention layers, while the remaining five are part of the overall comment attention and encoding layers.

The Output Shape column in tables 4.1 and 4.2 is a tuple that represents the size and shape of the given layer. The None in each tuple is returned by Keras [13] to represent a variable batch size, as this is specified not when the model is created but when it is trained. The remaining values provide the vector size as well as any other additional dimensions included in the layer. As the GloVe word embeddings used have a feature vector of size 100, this number is shown frequently within the architecture of the neural networks.

Layer (type)	Output Shape	Param #
Input	(None, 100)	0
Embedding	(None, 100, 100)	2466000
Bi-Directional	(None, 100, 200)	120600
Attention	(None, 200)	20200
Input	(None, 20, 100)	0
Time Distributed	(None, 20, 200)	2606800
Bi Directional	(None, 20, 200)	180600
Attention	(None, 200)	20200
Dense	(None, 2)	402

Table 4.1: Architecture for basic attention model, extending Yang et al. (2016)'s architecture . The two attention and embedding layers have been separated within the table.

Hyper parameter tuning was facilitated using the Hyperas [32] library, which allows for intelligent choice between multiple hyper-parameters. After including this, the architecture was expanded to the layers shown in Table 4.2, which included multiple layers with 2-3 choices for Hyperas to select. Hyperas was also able to choose between 5 and 10 epochs, a batch size of either 50, 64, or 128, and either Root Mean Square (RMS) Propagation,

[4] Code for the HAN was adapted from this implementation: `https://github.com/richliao/textClassifier/blob/master/textClassifierHATT.py`.

Adam Optimizer, or Stochastic Gradient Descent (SGD).

Layer (type)	Output Shape	Param #
Input	(None, 100)	0
Embedding	(None, 100, 100)	2466000
Bi-Directional	(None, 100, 200)	120600
Attention	(None, 200)	20200
Input	(None, 20, 100)	0
Time Distributed	(None, 20, 200)	2606800
Bi Directional	(None, 20, 200)	180600
Attention	(None, 200)	20200
Dense	(None, *256/512/1024*)	102912
Activation (*ReLU/Sigmoid*)	(None, 512)	0
Dropout (Uniform, *0/1*)	(None, 512)	0
Dense	(None, 10)	5130
Dense	(None, 2)	22

Table 4.2: Architecture for optimized attention model. Choices given to the hyper-parameter tuner are shown in *italics*. Parameter numbers may change depending on the model - these are shown for the best model selected by Hyperas.

4.2 Evaluation

The success of these classifiers was evaluated by several metrics such as accuracy, precision, recall, and F-measure (F_1). Hyper-parameter tuning of the models used Area under the Curve (AUC), and Receiving Operation Characteristic (ROC) curves [39] to evaluate performance. Results from scikit-learn include precision and recall for both classes, treating each label as positive and negative classes respectively.

Precision (P) and recall (R) are common metrics of effectiveness that can be extracted from a confusion matrix:

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

where TP represents the true-positive results (data points in which the algorithm correctly labels the positive class correctly), FP is the false-positive results (in which a negative data point is labeled incorrectly as positive), and FN is the false-negative results (a positive data point is incorrectly labeled as negative). The F_1 score is the harmonic mean of Precision and Recall, and can be expressed as the following.

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R}$$

The ROC curve is created by plotting the precision and fall-out against each other. Fallout, or false-positive rate, is defined as

$$FPR = \frac{FP}{FP - FN}$$

and represents the likelihood that a false positive will occur. From the ROC, one can derive the AUC as area under the ROC curve.

Micro, macro, and weighted averages are also included in scikit-learn's evaluation of models. These averages use the positive and negative results for precision and recall to average out the values, and are implemented by scikit-learn[5] by the following definitions:

- Micro average: averaging the total true positives, false negatives and false positives;
- Macro average: averaging the unweighted mean per label;
- Weighted average: averaging the support (N) weighted mean per label.

[5] Scikit-learn documentation for performance metrics can be viewed at: `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html`

The models were evaluated by the collected annotation data, which has been completely held-out from training, as well as the validation split of the overall dataset used for training. During the annotation task, the dataset shown to the annotators was rebalanced to contain 75% Removed comments and 25% Not Removed, as the study was primarily interested in understanding removal reasons. This distribution is visible in the subsets of comment annotations. Of the held-out annotation data, only the comments that have been annotated twice were used for evaluation. The following experiments were used to evaluate the performance of the models:

- (Experiment 1) Original tag (overall N = 8,284, val. N = 1,657)

  This experiment used the full database of comments. The majority class of this dataset was Not Removed, at roughly 75% Not Removed and 25% Removed for the training data, and a 76% and 24% split for the validation data.

- (E2) Original tag (overall N = 3,352, val. N = 670)

  This experiment used a label-balanced subset of the original database of comments, reducing the size by several thousand comments. As the dataset has been balanced, the distribution of the labels are 50% Removed and 50% Not Removed.

- (E3) Original tag (overall N = 2,235, val. N = 447)

  This experiment used an even smaller subset of the original database of comments, balanced such that the majority class is Removed, comprising of 75% of the database.

- (E4) Agreement - annotator tag (val. N = 171)

  For the twice-annotated comments where the two annotators agreed on a label, the consensus replaced the original label in this subset as the ground truth. The majority class was Removed, at 77.8%.

- (E5) Agreement - original tag (val. N = 171)

  For the twice-annotated comments where the two annotators agreed on a label, the original tag was used in this experiment instead of the consensus from annotators.

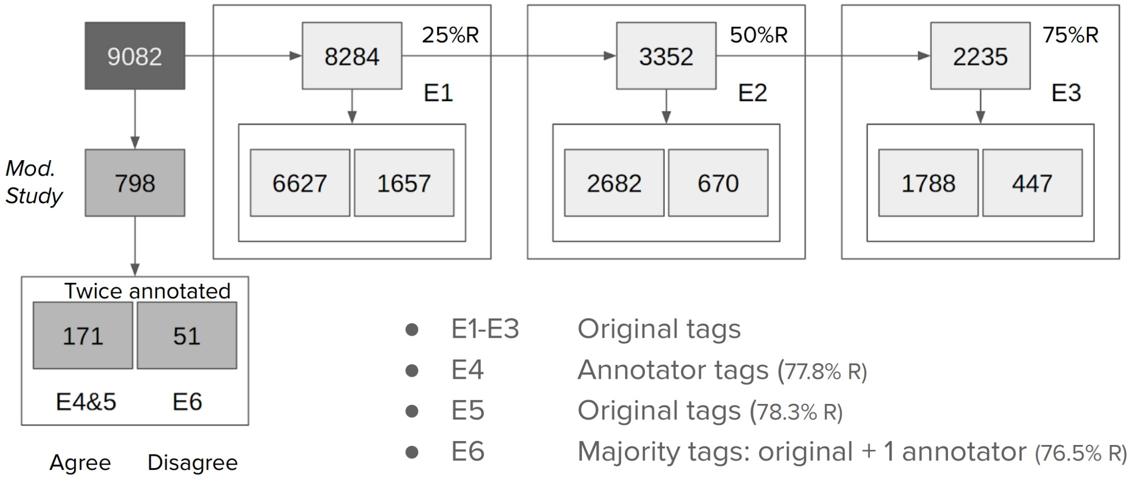


Figure 4.1: A visual representation of how data has been split and separated for the various experiments. The initial full set of 9,082 received 1,020 annotations over 798 unique comments, which were removed for E1-E3. Of these comments, 222 received two annotations, which were split into Agree for E4 and E5, and Disagree for E6. E1-E3 use the same set of Removed comments, with decreasings subset of Not Removed comments for E2 and E3 both training and validation sets.

The results from E2 and E3 therefore can jointly compare performance over both the annotator tags and original tags for data that was more straightforward for annotators to label. The majority class was Removed, at 78.3% of the comments in this subset.

- (E6) Disagreement - original tag (val. N = 51)

  The smallest subset at 51 comments, this included the twice-annotated comments for which the two annotators disagreed. As one of the two annotators must then agree with the original tag, the original tag served as the majority opinion between the three sources. The majority class for this dataset was also Removed, at 76.5% of the comments.

For each experiment E4 through E6, all models trained over the datasets used in E1 through E3 were evaluated. Figure 4.1 shows the flow of data from the initial dataset as it is separated for each experiment as described above.

4.3 Results across Experiments

E #	Experiment	N	GNB	DT	SVC	DNN	% Rm	# AM Rm	% AM Rm	% Overlap
E1a	Original tag (U)	1657	61	73	**78**	-	24	101	6	28
E1b	Original tag (W)	1657	-	66	50	**72**	24	101	6	28
E2	Original tag	670	55	54	**59**	54	50	96	14	27
E3a	Original tag (U)	447	53	72	75	-	75	106	24	31
E3b	Original tag (W)	447	-	72	**75**	65	75	106	24	31
E4a	Agree - annotator (E1)	171	57	50	**80**	35	77.8	44	26	28
E4b	Agree - annotator (E2)	171	69	56	**81**	48	77.8	44	26	28
E4c	Agree - annotator (E3)	171	50	72	**78**	57	77.8	44	26	28
E5a	Agree - original (E1)	171	53	48	**72**	35	78.3	44	26	33
E5b	Agree - original (E2)	171	65	55	**74**	51	78.3	44	26	33
E5c	Agree - original (E3)	171	50	73	**78**	57	78.3	44	26	33
E6a	Disagree - original (E1)	51	47	39	**59**	33	76.5	10	15	20
E6b	Disagree - original (E2)	51	**75**	53	65	51	76.5	10	15	20
E6c	Disagree - original (E3)	51	49	65	**78**	63	76.5	10	15	20

Table 4.3: Percent accuracy of models across categories in percent. (U) and (W) refer to the unweighted and weighted models, respectively. In this table, and in the remainder of chapter four when discussing results, *N* refers to the size of the validation set. The abbreviation *Rm* stands for Removed for all applicable columns, and *AM* for AutoMod. *% AM Rm* refers to the percentage of comments that were removed by AutoMod, while the *% Overlap* is the percentage of Removed comments that were removed by AutoMod.

In Table 4.3, the results are separated by E1 through E3, which focus on the accuracy of the models over validation sets of differently balanced subsets of the training data, and E4 through E6 which evaluates the E1 through E3 models over the modestly-sized validation sets comprised of twice-annotated comments from the annotation task. The number of comments for these categories is much smaller than the overall dataset, with 171 comments in the agree subset, and 51 in the disagree subset. The N column refers to the size of the validation set for each experiment, and each model in each part of a given experiment uses the same set, with the exception of the DNN which generates the train/test split differently.

The amount of comments that were removed by AutoModerator varies slightly based on the validation set used due to differences in sampling. However, when inspecting the full dataset of 8,284 comments, the numbers are similar both for the percentage of AutoMod Removed comments, at 6%, and 28% for the *overlap*, or the percentage of the Removed comments that were removed by AutoModerator.

As shown in the rest of the tables in this chapter, both the Not Removed and Removed classes were treated as both positive and negative for the purpose of precision and recall metrics. However, the results for the Removed class are more interesting for the applied purpose of creating an automated moderation system that could support human moderation efforts by detecting and flagging troublesome comments for review.

Preliminary results on E1 for the DNN and SVC showed a strong tendency for them to label everything as Not Removed, and adjusting the models by weighing the classes was necessary to address this issue. In this context, class weighing refers to assigning weights to the classes based on their distributions. The less represented a given class is within the set, the heavier it is weighted and the more models are punished for labeling a member of a minority class incorrectly. Due to the increased time to train the DNN compared to the SVC, an unweighted DNN was not included for the experiments. The unweighted DT and SVC have been included for E1, E2 and E3 for comparison. As GNB was not weighted, results are not available for the GNB for weighted experiments.

4.4 E1 - Original Tag: 75% Not Removed

4.4.1 Unweighted Models

The confusion matrices in Figure 4.2 show the results across the unweighted models for E1. Of the unweighted models, the SVC had the highest accuracy overall accurcay of 78%, however this was attained by selecting the majority class, which is Not Removed, for all comments. This resulted in an F1 score of zero for the Remove class. As the SVC only selected Not Remove, Table 4.7 shows scores of zero for the precision, recall, and F1 for the Remove label.

The GNB and DT also performed better on the Not Removed class than the Removed,

	% Accuracy	Rm F1 Score
GNB	61	**0.31**
DT (U)	73	0.24
SVC (U)	**78**	0.00
DT (W)	66	0.39
SVC (W)	50	0.39
DNN	**72**	0.36

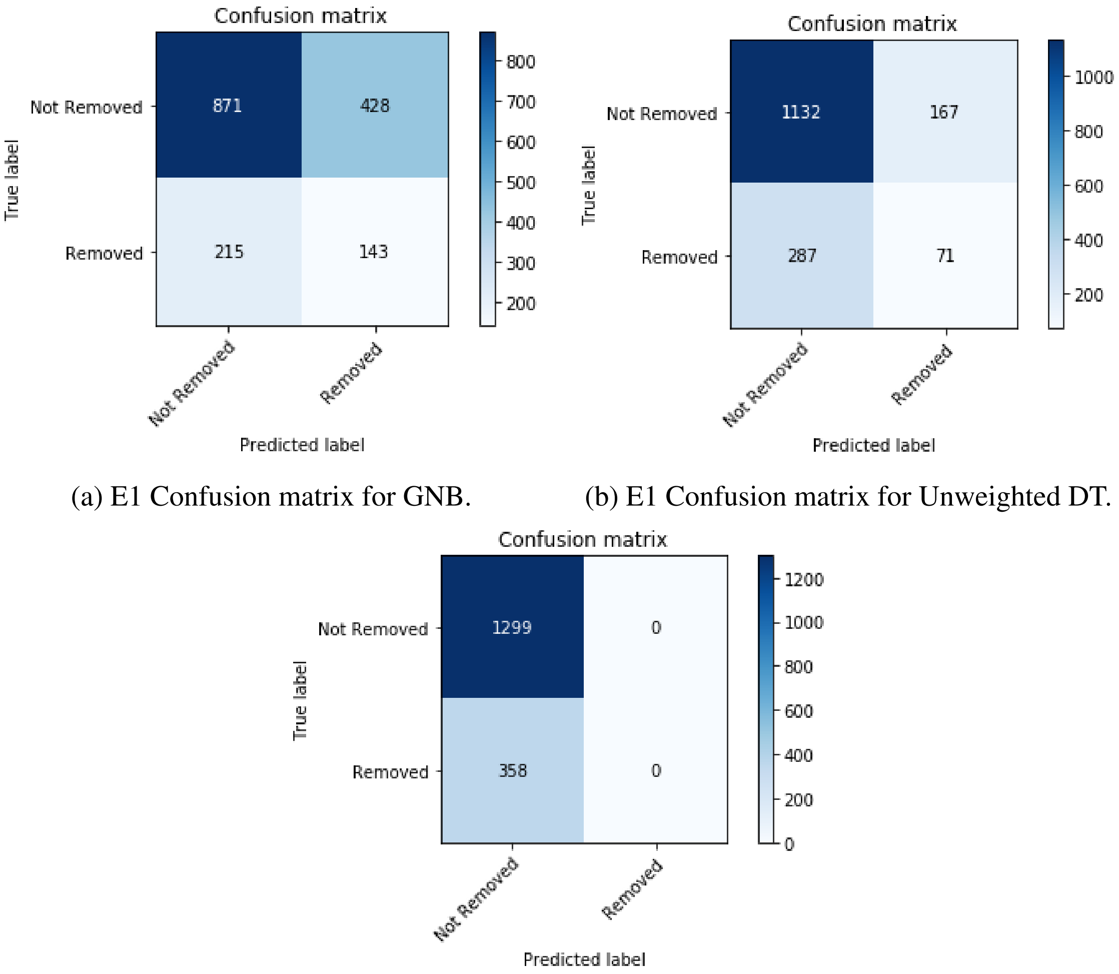
Table 4.4: Accuracy and F1 score for Removed class for all E1 models (N = 1657).

but were still able to correctly identify some Removed comments. However, the precision and recall for the Removed label was low at 0.25 and 0.40 for the GNB and 0.30 and 0.20 for the DT respectively, as shown in Tables 4.5 and 4.6. The GNB had a large amount of false positives, listing many comments as Removed that were Not Removed, and the DT instead had a higher instance of false negatives. However, the DT was more accurate, at 73% overall accuracy to the GNB's 61%.

4.4.2 Weighted Models

The weighted DT and SVC corrected the bias against the Removed class caused by the unbalanced data, and while the overall accuracy for the models suffered (66% and 50%, respectively), the scores for the Not Removed class improved substantially. As illustrated in Figure 4.3b, the weighted SVC had a strong tendency for false positives, labeling many Not Removed comments as Removed.

In comparison, the weighted DNN model chosen by Hypears performed at an accuracy of 72%, better than any of the weighted traditional models. Table 4.10 and Figure 4.3c show that the model scored better for the Not Removed class than the Removed class, which was also the case for the traditional models, implying that weighing the classes cannot altogether mitigate the effects of an unbalanced dataset.



(a) E1 Confusion matrix for GNB.

(b) E1 Confusion matrix for Unweighted DT.

(c) E1 Confusion matrix for Unweighted SVC.

Figure 4.2: E1 Confusion matrices for unweighted models (N = 1657).

	Precision	Recall	F-1 Score	N
Not removed	0.80	0.67	0.73	1299
Removed	0.25	0.40	0.31	358
Micro avg	0.61	0.61	0.61	1657
Macro avg	0.53	0.53	0.52	1657
Weighted avg	0.68	0.61	0.64	1657

Table 4.5: E1 GNB classification results (N = 1657).

	Precision	Recall	F-1 Score	N
Not removed	0.80	0.87	0.83	1299
Removed	0.30	0.20	0.24	358
Micro avg	0.73	0.73	0.73	1657
Macro avg	0.55	0.53	0.54	1657
Weighted avg	0.69	0.73	0.70	1657

Table 4.6: E1 Unweighted DT classification results (N = 1657).

	Precision	Recall	F-1 Score	N
Not removed	0.78	1.00	0.88	1299
Removed	0.00	0.00	0.00	358
Micro avg	0.78	0.78	0.78	1657
Macro avg	0.39	0.50	0.44	1657
Weighted avg	0.61	0.78	0.69	1657

Table 4.7: E1 Unweighted SVC classification results (N = 1657).

	Precision	Recall	F-1 Score	N
Not removed	0.84	0.70	0.76	1299
Removed	0.32	0.51	0.39	358
Micro avg	0.66	0.66	0.66	1657
Macro avg	0.58	0.60	0.58	1657
Weighted avg	0.73	0.66	0.68	1657

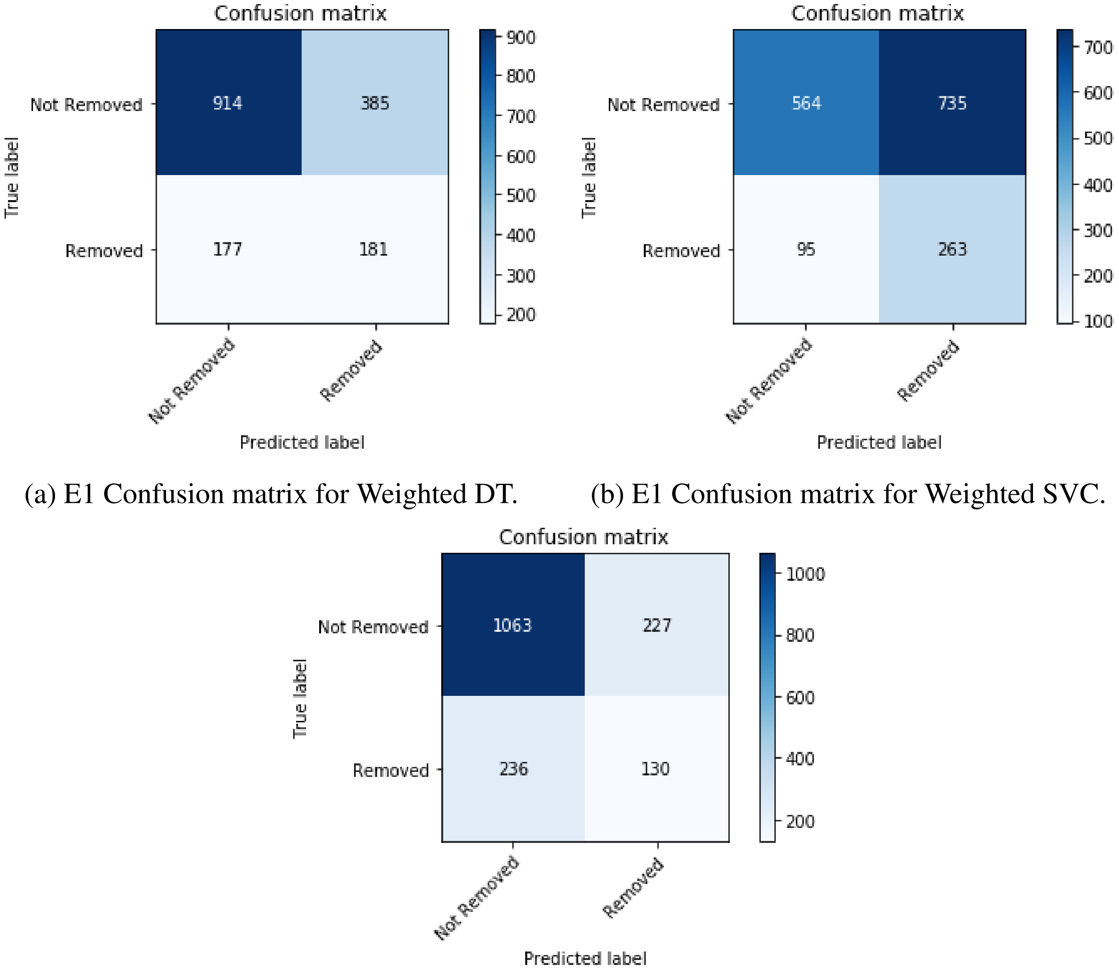
Table 4.8: E1 Weighted DT classification results (N = 1657).

	Precision	Recall	F-1 Score	N
Not removed	0.86	0.43	0.58	1299
Removed	0.26	0.73	0.39	358
Micro avg	0.50	0.50	0.50	1657
Macro avg	0.56	0.58	0.48	1657
Weighted avg	0.73	0.50	0.54	1657

Table 4.9: E1 Weighted SVC classification results (N = 1657).

	Precision	Recall	F-1 Score	N
Not removed	0.82	0.82	0.82	1290
Removed	0.36	0.36	0.36	366
Micro avg	0.72	0.72	0.72	1656
Macro avg	0.59	0.59	0.59	1656
Weighted avg	0.72	0.72	0.72	1656

Table 4.10: E1 Weighted DNN classification results (N = 1656).



(a) E1 Confusion matrix for Weighted DT.

(b) E1 Confusion matrix for Weighted SVC.

(c) E1 Confusion matrix for Weighted DNN (N = 1656).

Figure 4.3: E1 Confusion matrices for weighted models (N = 1657). (The N for DNN is 1656 due to a difference in rounding when splitting the test and training sets.)

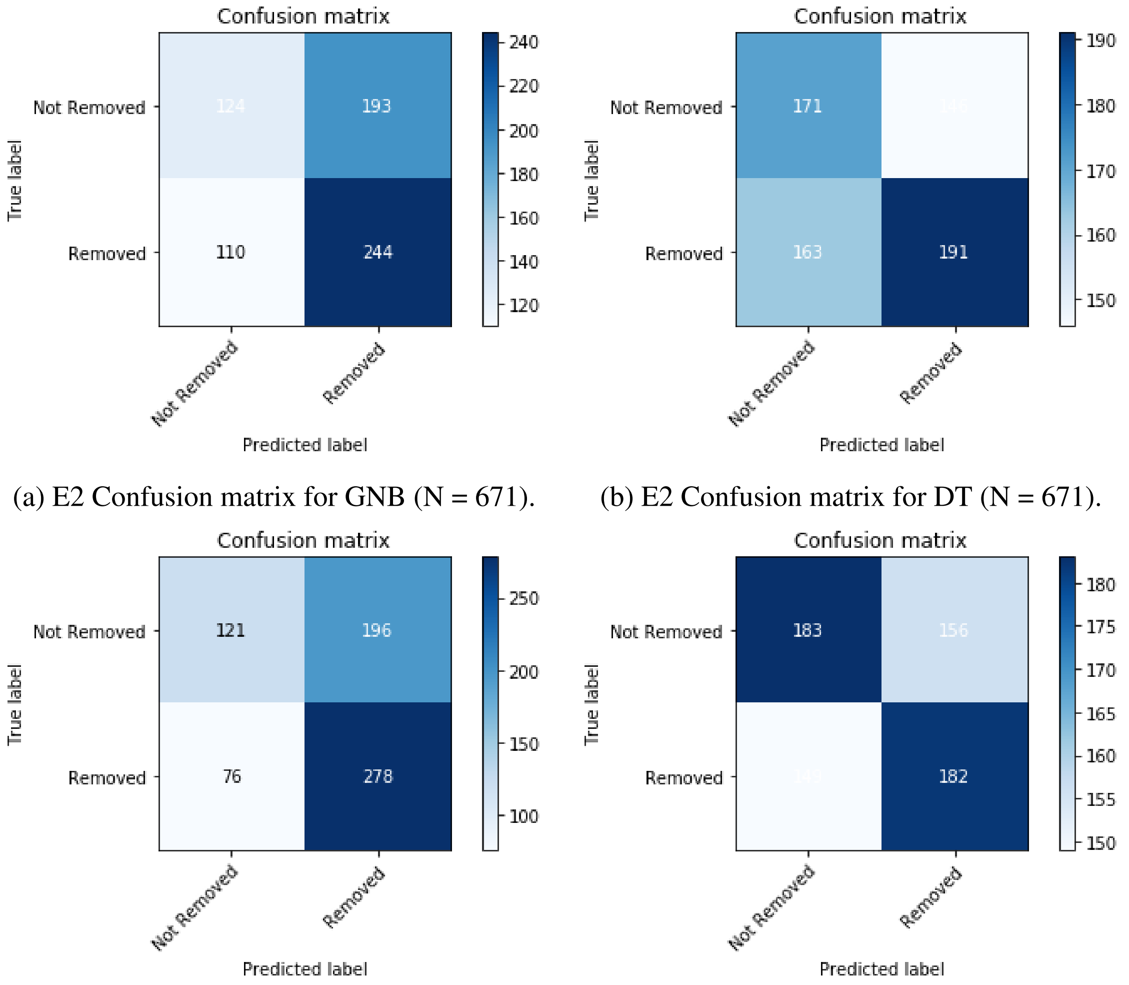
4.5 E2 - Original Tag: Balanced

By using a balanced dataset for E2, we hope to gain a better understanding of how the models perform without the additional bias of an unbalanced dataset. The SVC and GNB models for this experiment had a stronger tendency for false positives, with many comments mislabeled as Removed. However, the DT model tended towards false negatives for its classification errors. Despite these differences the GNB and DT had similar accuracies of 55% and 54%, highlighting that multiple metrics are needed to fully understand a model as the results for the two models appear visually different in Figures 4.4a and 4.4b.

	% Accuracy	Rm F1 Score
GNB	55	0.62
DT	54	0.55
SVC	**59**	**0.67**
DNN	54	0.54

Table 4.11: Accuracy and F1 score for Removed class for all E2 models (N = 670).

The DNN had similar results as E1 in that a similar amount of comments for both classes were misclassified, which can be seen in Table 4.4d. It was outperformed by the traditional models in this experiment.



(a) E2 Confusion matrix for GNB (N = 671). (b) E2 Confusion matrix for DT (N = 671).

(c) E2 Confusion matrix for SVC (N = 671). (d) E2 Confusion matrix for DNN (N = 670).

Figure 4.4: E2 Confusion matrices - original tag results. (The N of the validation set for the traditional models was 671 and 670 for the DNN, due to a difference in rounding.)

	Precision	Recall	F-1 Score	N
Not removed	0.53	0.39	0.45	317
Removed	0.56	0.69	0.62	354
Micro avg	0.55	0.55	0.55	671
Macro avg	0.54	0.54	0.53	671
Weighted avg	0.54	0.55	0.54	671

Table 4.12: E2 GNB classification results (N = 671).

	Precision	Recall	F-1 Score	N
Not removed	0.51	0.54	0.53	317
Removed	0.57	0.54	0.55	354
Micro avg	0.54	0.54	0.54	617
Macro avg	0.54	0.54	0.54	671
Weighted avg	0.54	0.54	0.54	671

Table 4.13: E2 DT classification results (N = 671).

	Precision	Recall	F-1 Score	N
Not removed	0.61	0.38	0.47	317
Removed	0.59	0.79	0.67	354
Micro avg	0.59	0.59	0.59	671
Macro avg	0.60	0.58	0.57	671
Weighted avg	0.60	0.59	0.58	671

Table 4.14: E2 SVC classification results (N = 671).

	Precision	Recall	F-1 Score	N
Not removed	0.55	0.54	0.55	339
Removed	0.54	0.55	0.54	331
Micro avg	0.54	0.54	0.54	670
Macro avg	0.54	0.54	0.54	670
Weighted avg	0.54	0.54	0.54	670

Table 4.15: E2 DNN classification results (N = 670).

4.6 E3 - Original Tag: 75% Removed

This subset was balanced so that it contains 75% Removed comments and 25% Not Removed comments, at the cost of a much smaller sample size. The validation set for this experiment contains only 447 comments.

	% Accuracy	Rm F1 Score
GNB	53	0.65
DT (U)	72	0.82
SVC (U)	75	**0.86**
DT (W)	72	0.83
SVC (W)	**75**	0.85
DNN	65	0.76

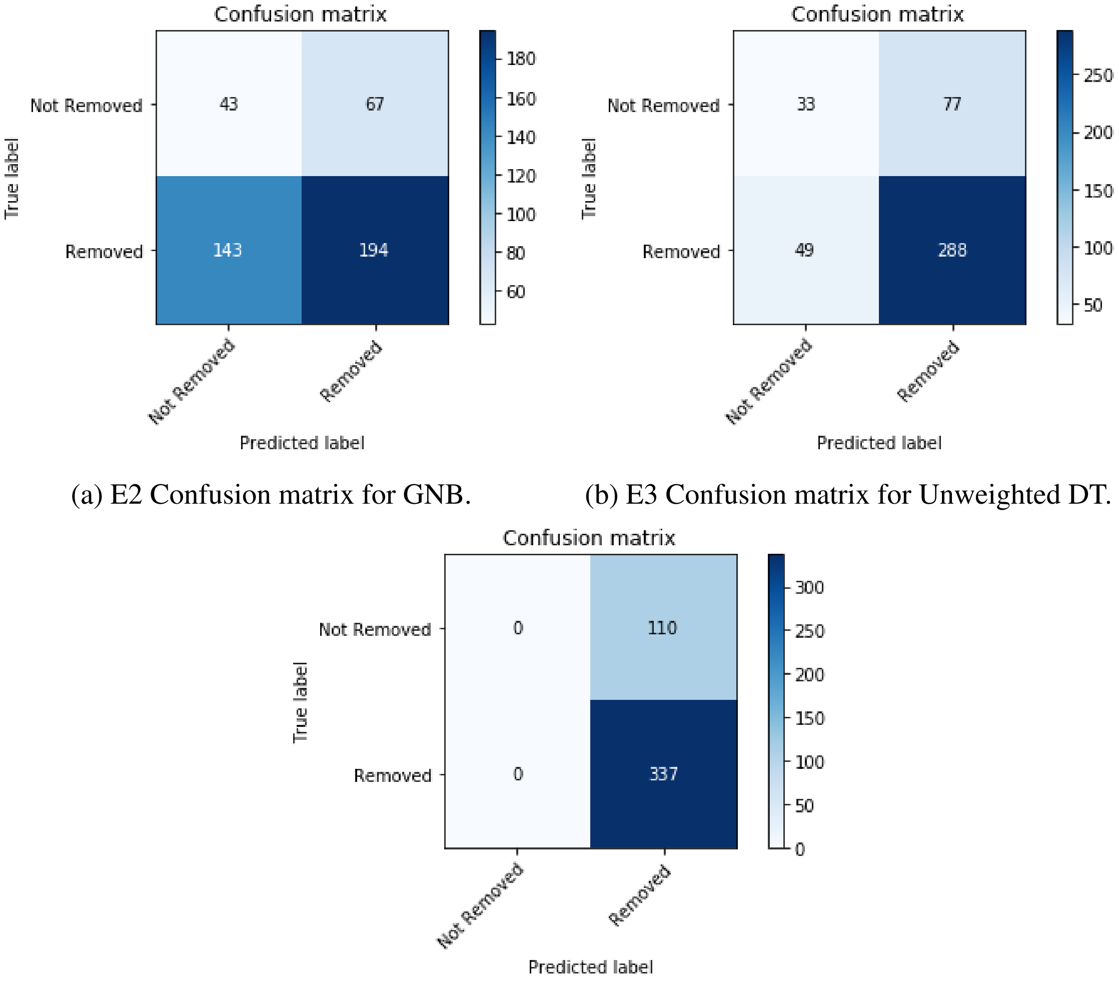
Table 4.16: Accuracy and F1 score for Removed class for all E3 models (N = 447).

4.6.1 Unweighted Models

Of the unweighted models, the GNB had the lowest accuracy of 53%. The confusion matrix in Figure 4.5a shows a large number of false negatives, in which the model was more likely to assign Not Removed to Removed comments. This is reflected in the low precision and recall scores for the Not Removed class, and the lower recall score for the Removed class shown in Table 4.17.

The SVC model achieved the highest accuracy of the group at 75% by labeling all comments as Removed, which can be seen clearly in Figure 4.5c, and reflected in Table 4.19 with a zeroed row for the Not Removed class.

While a lower overall accuracy than the SVC, the DT model has a similar accuracy score of 72%, and successfully labels several Not Removed comments. The confusion matrix in 4.5b shows a more even distribution of false positives and negatives, with less comments correctly labeled Not Removed than the GNB.



(a) E2 Confusion matrix for GNB.

(b) E3 Confusion matrix for Unweighted DT.

(c) E3 Confusion matrix for Unweighted SVC.

Figure 4.5: E3 Confusion matrices for unweighted models (N = 447).

	Precision	Recall	F-1 Score	N
Not removed	0.23	0.39	0.29	110
Removed	0.74	0.58	0.65	337
Micro avg	0.53	0.53	0.53	447
Macro avg	0.49	0.48	0.47	447
Weighted avg	0.62	0.53	0.56	447

Table 4.17: E3 GNB classification results (N = 447).

	Precision	Recall	F-1 Score	N
Not removed	0.40	0.30	0.34	110
Removed	0.79	0.85	0.82	337
Micro avg	0.72	0.72	0.72	447
Macro avg	0.60	0.58	0.58	447
Weighted avg	0.69	0.72	0.70	447

Table 4.18: E3 Unweighted DT classification results (N = 477).

	Precision	Recall	F-1 Score	N
Not removed	0.00	0.00	0.00	110
Removed	0.75	1.00	0.86	354
Micro avg	0.75	0.75	0.75	477
Macro avg	0.38	0.50	0.43	477
Weighted avg	0.57	0.75	0.65	477

Table 4.19: E3 Unweighted SVC classification results (N = 477).

4.6.2 Weighted Models

	Precision	Recall	F-1 Score	N
Not removed	0.29	0.11	0.16	110
Removed	0.76	0.91	0.83	337
Micro avg	0.72	0.72	0.72	447
Macro avg	0.53	0.51	0.49	447
Weighted avg	0.64	0.72	0.66	447

Table 4.20: E3 Weighted DT classification results (N = 477).

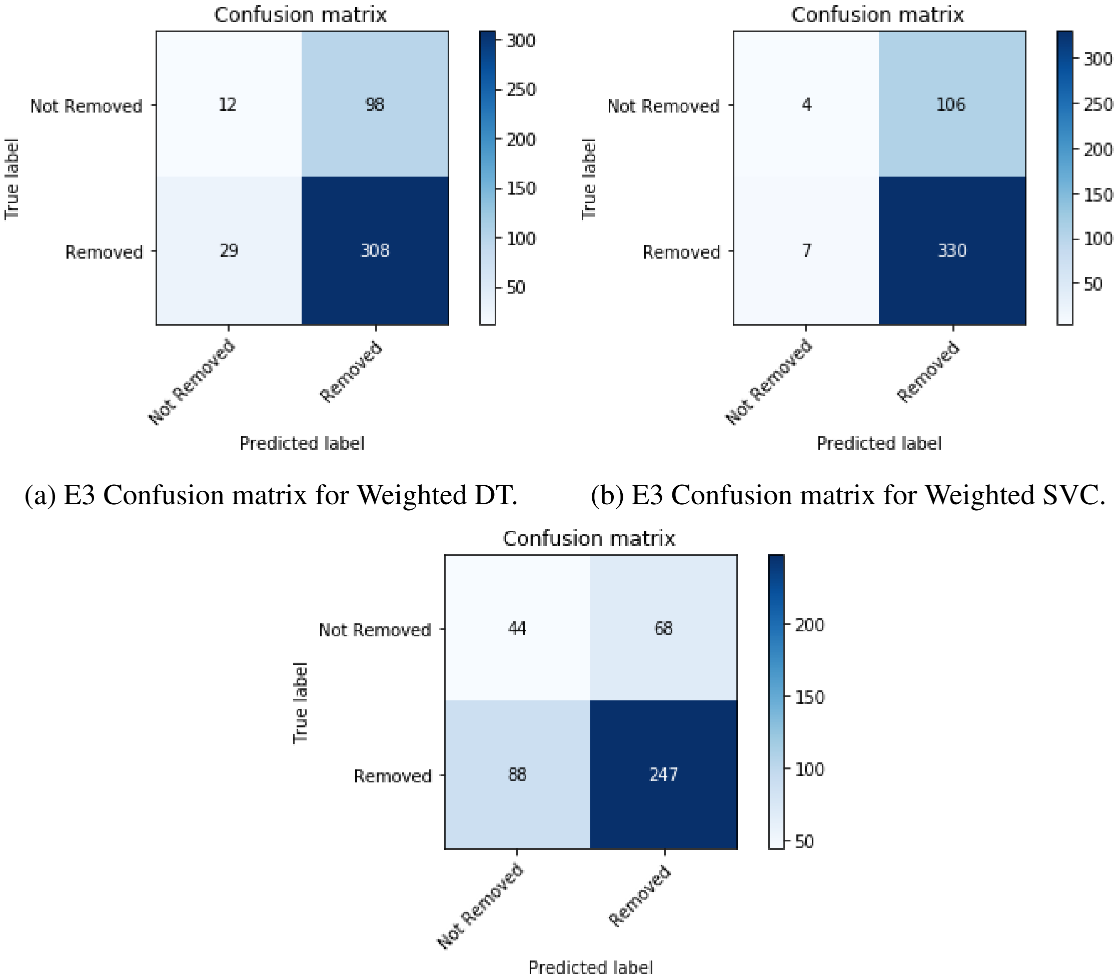
	Precision	Recall	F-1 Score	N
Not removed	0.36	0.04	0.07	110
Removed	0.76	0.98	0.85	354
Micro avg	0.75	0.75	0.75	477
Macro avg	0.56	0.51	0.46	477
Weighted avg	0.66	0.75	0.66	477

Table 4.21: E3 Weighted SVC classification results (N = 477).

	Precision	Recall	F-1 Score	N
Not removed	0.33	0.39	0.36	112
Removed	0.78	0.74	0.76	335
Micro avg	0.65	0.65	0.65	447
Macro avg	0.56	0.57	0.56	447
Weighted avg	0.67	0.65	0.66	447

Table 4.22: E3 Weighted DNN classification results (N = 447).

The DNN performed similarly to the weighted traditional models over the E3 validation set with an accuracy of 65%, and had the highest F1 score for the Not Removed class of any of the models, as shown in in Table 4.22. Figure 4.6c shows that the model misclassified a similar number of comments for both labels, resulting in a higher percentage of mislabeled Not Removed comments.



(a) E3 Confusion matrix for Weighted DT.

(b) E3 Confusion matrix for Weighted SVC.

(c) E3 Confusion matrix for Weighted DNN.

Figure 4.6: E3 Confusion matrices for weighted models (N = 447).

Interestingly, the weighted DT had similar if not worse results than the unweighted DT, but with the same overall accuracy of 72%. Inspecting the precision and recall in Table 4.20, a reduction in both metrics for the Not Removed class is visible. The weighted DT presented a much heavier bias toward labelling comments as Removed, and a higher amount of false positives for the Removed class.

Unlike the weighted SVC in E1, the weighted SVC for E3 did not seem to improve from weighing the classes, and were more sensitive to the class imbalance in the data. This may be due to an issue in the generation of the model, or chance that the GridSearchCV optimizer chose a more balanced model for E1 and a model heavily biased towards the majority class for E3. As shown in 4.6b, eleven comments were labeled as Not Removed out of the validation set. The recall value of 0.04, shown in Table 4.21, for the Not Removed class also reflects this bias in the SVC.

4.7 E4 - Agree (Annotator Tag)

The original tags were replaced with the consensus from the annotators for the label. This changes the distribution from the Agree (original tag) set slightly against Not Removed, at a distribution of 77.8% (from 78.3) Removed and 22.2% (from 21.7) Not Removed.

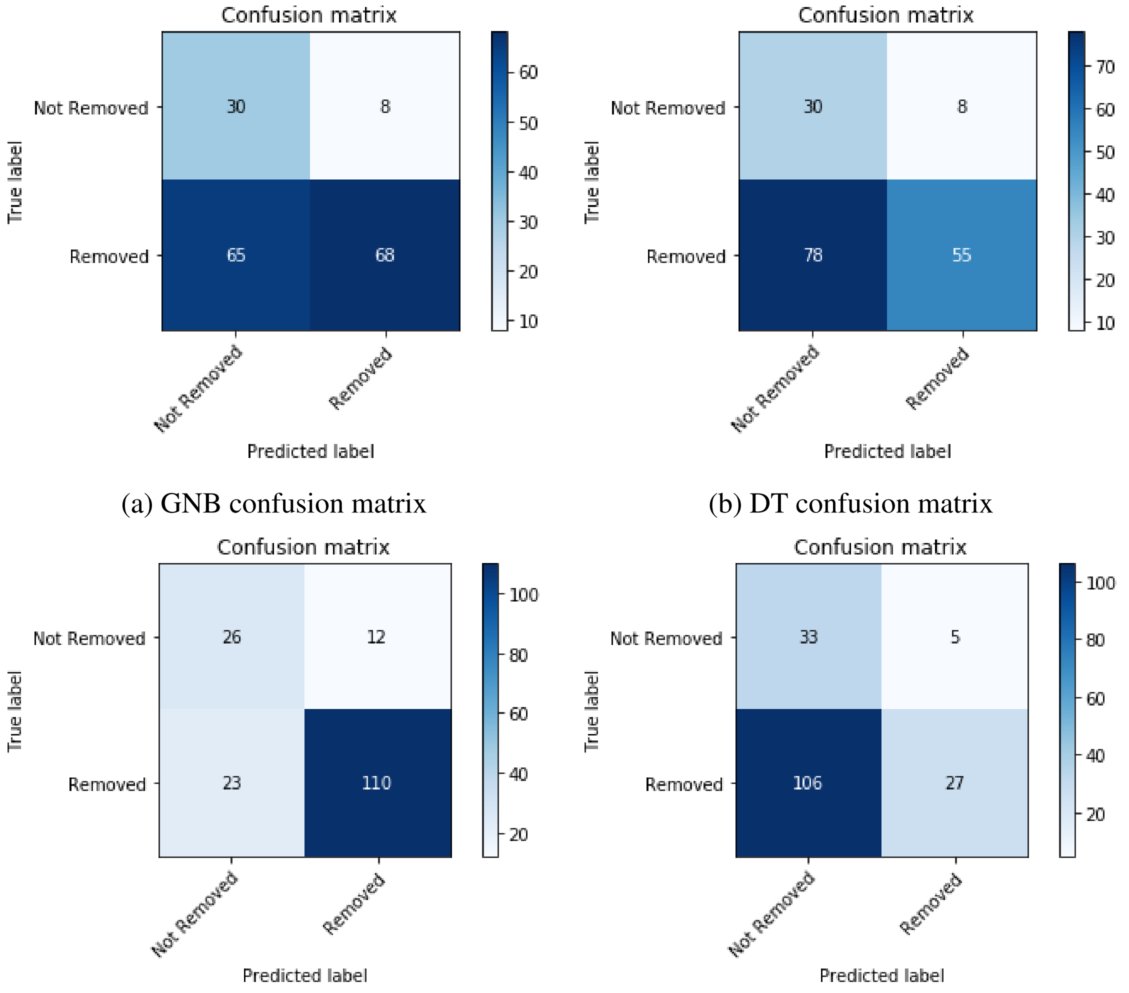
	GNB	DT	SVC	DNN
E1	57	50	**80**	35
E2	69	56	**81**	48
E3	50	72	**78**	57

Table 4.23: Percent accuracy for weighted E1, E2, and E3 models across E4 set.

As shown in Table 4.23, the SVC outperformed all other models for each experiment. Interestingly, the accuracy for each experiment was roughly equal for the SVC, despite the different distributions of data. The SVC trained on E3 chooses the majority class, but the E1 and E2 SVCs seem to be more resilient to the challenges presented by using the held-out annotation data. This could be due to the nature of the model; SVCs in general are less susceptible to overfitting than DT models.

4.7.1 E1 Model Performance

The GNB and weighted DT, SVC, and DNN from E1 were evaluated over this subset of comments. All traditional models saw a variation in accuracy from E1, a reduction from 66% to 50% for DT, an increase from 50% to 80% for SVC, and a decrease from 61% to 57% for GNB. However, the DNN experienced a severe drop in performance, lowering from 72% accuracy to 35%.



(a) GNB confusion matrix

(b) DT confusion matrix

(c) SVC confusion matrix

(d) DNN confusion matrix

Figure 4.7: E4 confusion matrices for E1 models (N = 171).

This reduction in performance ror all models, excluding the SVC, for this subset can be explained by the models not being trained on any annotator labels, as these were held out for evaluation. Additional training of the models using the annotation labels may help mitigate this issue. Under the scenario of adapting these models to a flagging system on Reddit, false positives for the Removed class are less important than false negatives. The SVC showed a much lower tendency to create false negatives than the other models, while also having a high classification rate for the Removed class.

	Precision	Recall	F-1 Score	N
Not removed	0.32	0.79	0.45	38
Removed	0.89	0.51	0.65	133
Micro avg	0.57	0.57	0.57	171
Macro avg	0.61	0.65	0.55	171
Weighted avg	0.77	0.57	0.61	171

Table 4.24: E4 GNB Classification result, using E1 model (N = 171).

	Precision	Recall	F-1 Score	N
Not removed	0.28	0.79	0.41	38
Removed	0.87	0.41	0.56	133
Micro avg	0.50	0.50	0.50	171
Macro avg	0.58	0.60	0.49	171
Weighted avg	0.74	0.50	0.53	171

Table 4.25: E4 DT Classification result, using E1 model (N = 171).

	Precision	Recall	F-1 Score	N
Not removed	0.53	0.68	0.60	38
Removed	0.90	0.83	0.86	133
Micro avg	0.80	0.80	0.80	171
Macro avg	0.72	0.76	0.73	171
Weighted avg	0.82	0.80	0.80	171

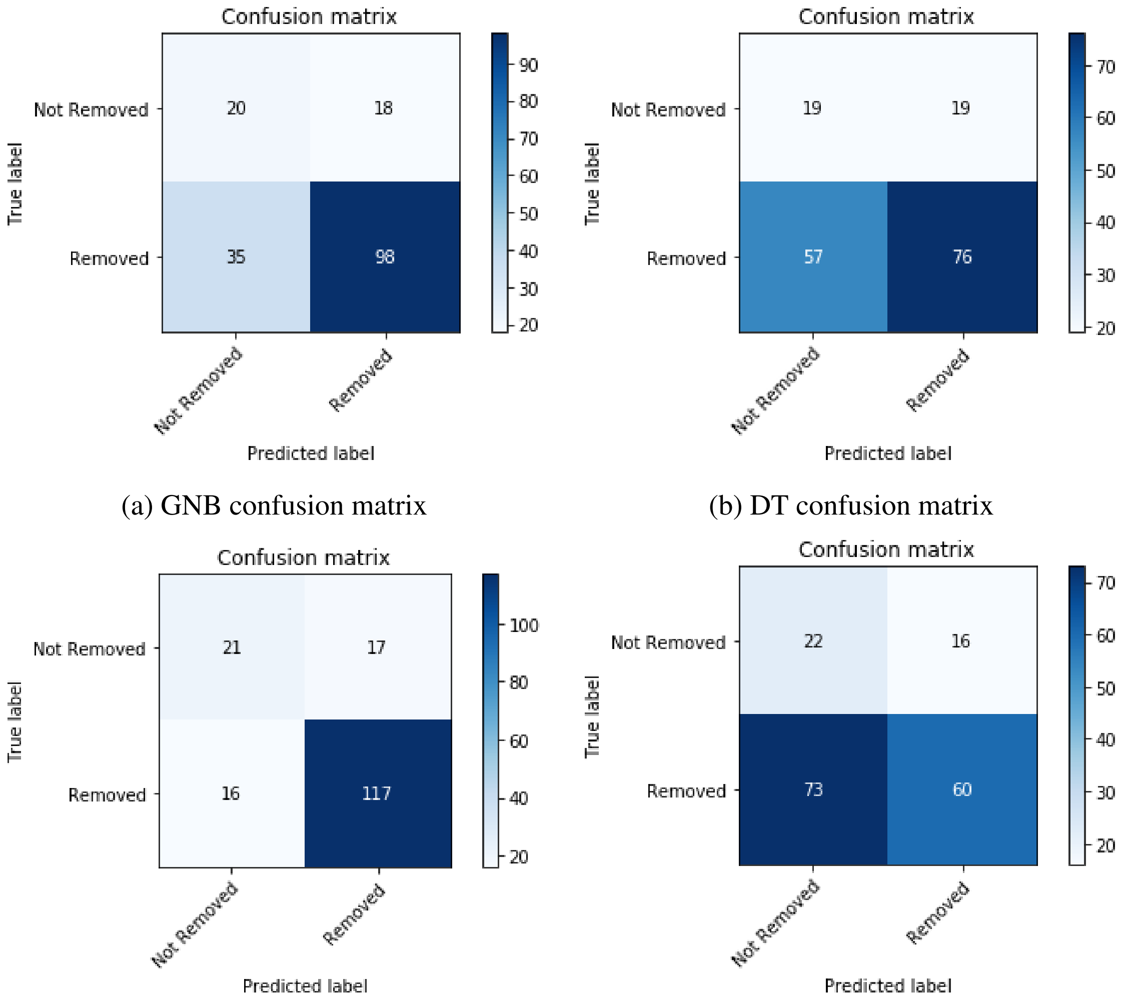
Table 4.26: E4 SVC Classification result, using E1 model (N = 171).

	Precision	Recall	F-1 Score	N
Not removed	0.24	0.87	0.37	38
Removed	0.84	0.20	0.33	133
Micro avg	0.35	0.35	0.35	171
Macro avg	0.54	0.54	0.35	171
Weighted avg	0.71	0.35	0.34	171

Table 4.27: E4 DNN Classification result, using E1 model (N = 171).

4.7.2 E2 Model Performance

Unsurprisingly, the SVC did extremely well on this subset, at 81% accuracy, in comparison to the E2 results. The E2 SVC model had a strong tendency toward false positives for the Removed class, which can be seen in 4.4c. A detriment on the E2 validation set, this tendency leads to a higher accuracy for E4 through E6 due to the unbalanced nature of the annotation data. The DNN again suffered from a reduction in accuracy, lowering from 54% on the E2 validation set to 48% for the E4 annotation set. The DT and DNN had a similar propensity for false negatives, as visible in Figure 4.8.



(a) GNB confusion matrix

(b) DT confusion matrix

(c) SVC confusion matrix

(d) DNN confusion matrix

Figure 4.8: E4 confusion matrices for E2 models (N = 171).

	Precision	Recall	F-1 Score	N
Not removed	0.36	0.53	0.43	38
Removed	0.84	0.74	0.79	133
Micro avg	0.69	0.69	0.69	171
Macro avg	0.60	0.63	0.61	171
Weighted avg	0.74	0.69	0.71	171

Table 4.28: E4 GNB Classification result, using E2 model (N = 171).

	Precision	Recall	F-1 Score	N
Not removed	0.25	0.50	0.33	38
Removed	0.80	0.57	0.67	133
Micro avg	0.56	0.56	0.56	171
Macro avg	0.53	0.54	0.50	171
Weighted avg	0.68	0.56	0.59	171

Table 4.29: E4 DT Classification result, using E2 model (N = 171).

	Precision	Recall	F-1 Score	N
Not removed	0.57	0.55	0.56	38
Removed	0.87	0.88	0.88	133
Micro avg	0.81	0.81	0.81	171
Macro avg	0.72	0.72	0.72	171
Weighted avg	0.81	0.81	0.81	171

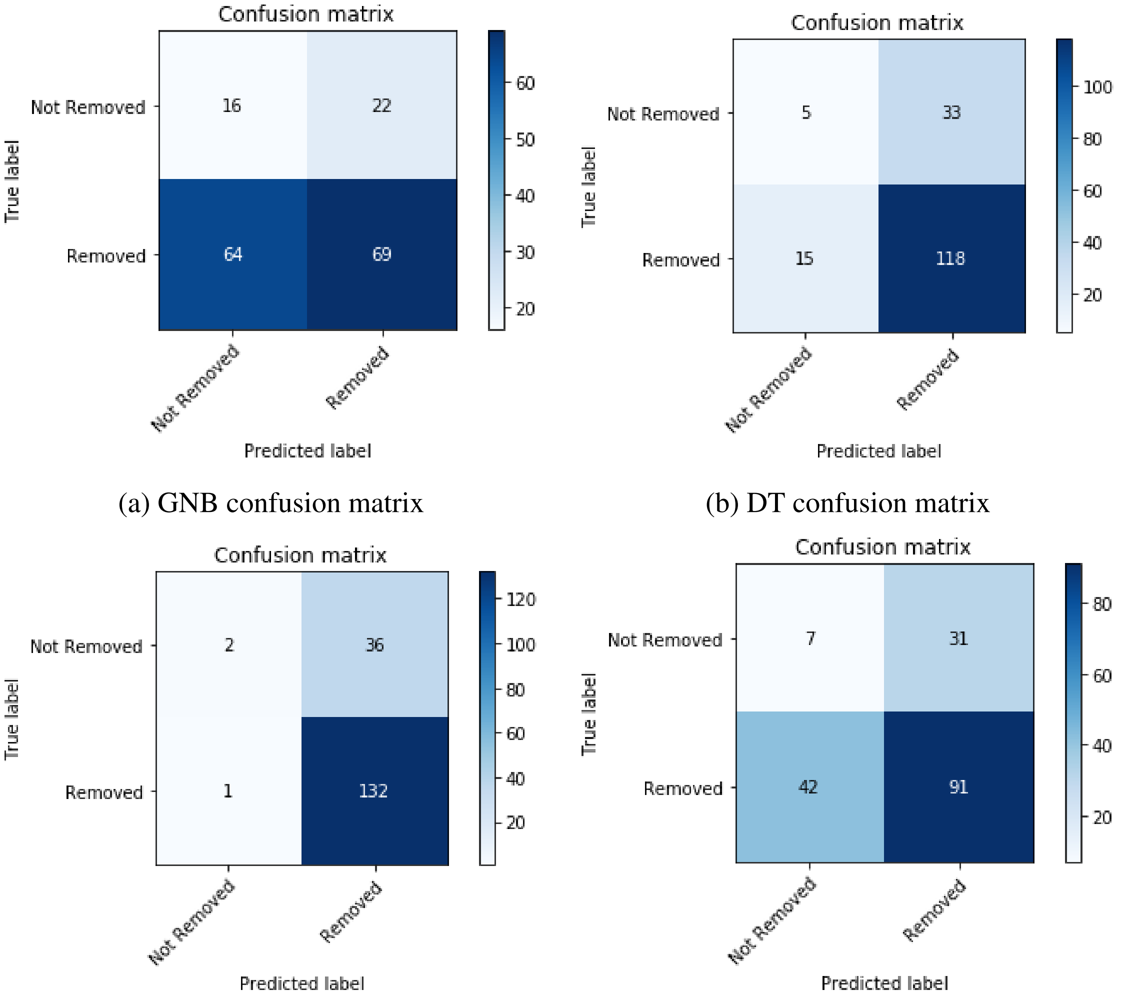
Table 4.30: E4 SVC Classification result, using E2 model (N = 171).

	Precision	Recall	F-1 Score	N
Not removed	0.23	0.58	0.33	38
Removed	0.79	0.45	0.57	133
Micro avg	0.48	0.48	0.48	171
Macro avg	0.51	0.52	0.45	171
Weighted avg	0.67	0.48	0.52	171

Table 4.31: E4 DNN Classification result, using E2 model (N = 171).

4.7.3 E3 Model Performance

Given that the SVC and DT E3 models tended to select the majority class shown in the previous section, these models performed well on this subset at 78% and 72% respectively. The E3 DNN also experienced a decrease in accuracy, from 65% on the E3 validation set to 57% for E4. The GNB maintained a similar level of performance, slightly lowered from 53% accuracy in E3 to 50%. The confusion matrices in Figure 4.9 show that the GNB and DNN both tend toward false negatives, where the SVC and DT labeled almost all comments as Removed.



(a) GNB confusion matrix

(b) DT confusion matrix

(c) SVC confusion matrix

(d) DNN confusion matrix

Figure 4.9: E4 confusion matrices for E3 models (N = 171).

	Precision	Recall	F-1 Score	N
Not removed	0.20	0.42	0.27	38
Removed	0.76	0.52	0.62	133
Micro avg	0.50	0.50	0.50	171
Macro avg	0.48	0.47	0.44	171
Weighted avg	0.63	0.50	0.54	171

Table 4.32: E4 GNB Classification result, using E3 model (N = 171).

	Precision	Recall	F-1 Score	N
Not removed	0.25	0.13	0.17	38
Removed	0.78	0.89	0.83	133
Micro avg	0.72	0.72	0.72	171
Macro avg	0.52	0.51	0.50	171
Weighted avg	0.66	0.72	0.68	171

Table 4.33: E4 DT Classification result, using E3 weighted model (N = 171).

	Precision	Recall	F-1 Score	N
Not removed	0.67	0.05	0.10	38
Removed	0.79	0.99	0.88	133
Micro avg	0.78	0.78	0.78	171
Macro avg	0.73	0.52	0.49	171
Weighted avg	0.76	0.78	0.70	171

Table 4.34: E4 SVC Classification result, using E3 weighted model (N = 171).

	Precision	Recall	F-1 Score	N
Not removed	0.14	0.18	0.16	38
Removed	0.75	0.68	0.71	133
Micro avg	0.57	0.57	0.57	171
Macro avg	0.44	0.43	0.44	171
Weighted avg	0.61	0.57	0.59	171

Table 4.35: E4 DNN Classification result, using E3 model (N = 171).

4.8 E5 - Agree (Original Tag)

The E5 validation subset includes the original tags for comments in which the annotators agreed, with a distribution of 78.3% Removed and 21.7% Not Removed. All models performed better on E4 than this subset, suggesting that the majority label between the two annotators and the original tag is easier for the models to classify than the original tags in the dataset.

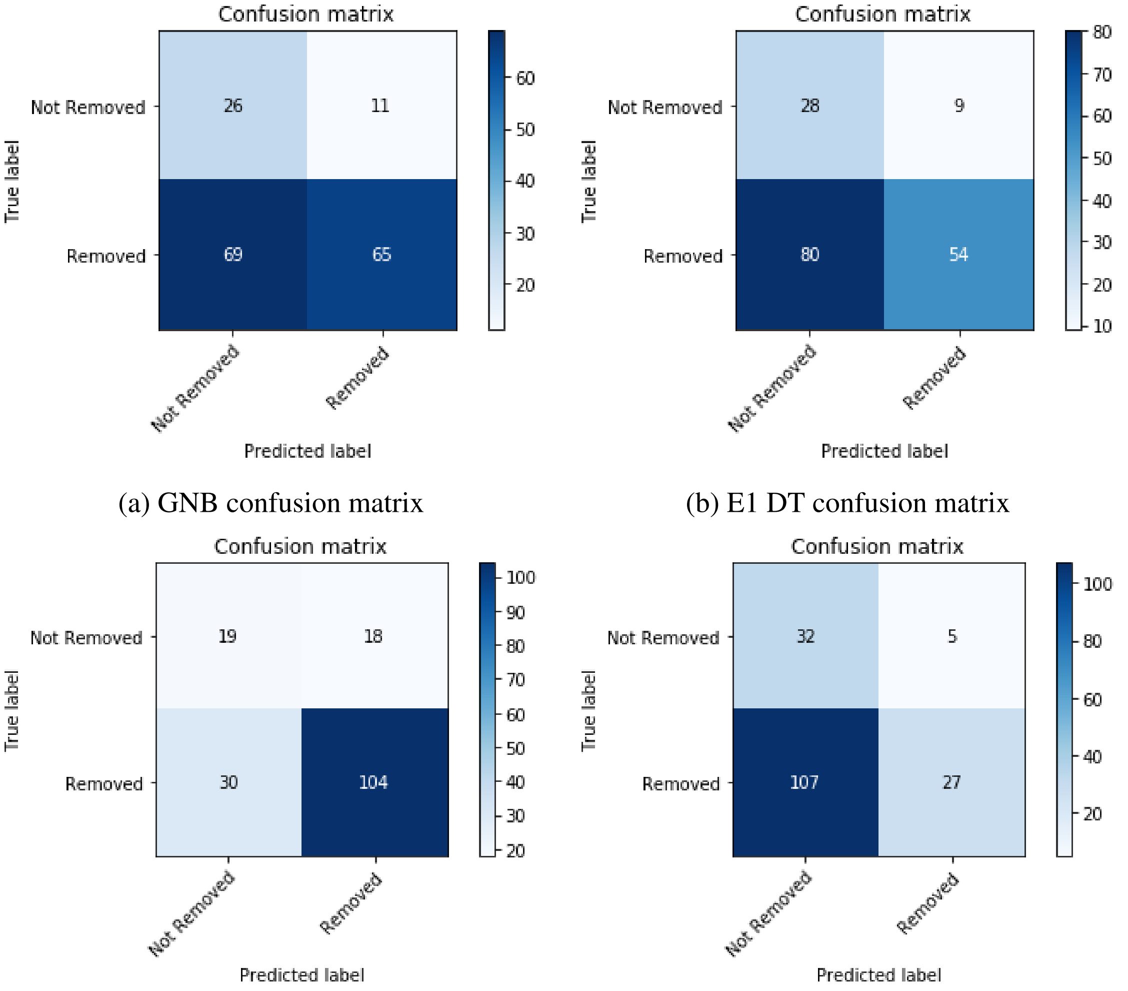
	GNB	DT	SVC	DNN
E1	53	48	**72**	35
E2	65	55	**74**	51
E3	50	73	**78**	57

Table 4.36: Percent accuracy for E1, E2, and E3 models for the E5 validation set.

Many of the E1 and E2 models experienced a drop in accuracy in their performance on the E5 set as compared to their results for E4. However, the results for the E3 models in terms of accuracy remained similar, which can be seen in Table 4.36. This is most likely due to the nature of the E3 models to mostly select Removed comments, notably the SVC and DT which almost exclusively pick the Removed class for E3. Once again the SVC outperformed all other models, with the E3 DT having a similarly high result.

4.8.1 E1 Model Performance

The GNB and DT models had an accuracy of 53% and 48% respectively, and all models tended toward false negatives for the Removed class for their misclassification errors. The DNN in particular struggled with this, with a slight decrease in accuracy from 31% in E4 to 30%, and all but one misclassification a false negative. The SVC remained the most accurate for the E5 set, but decreased from 80% in E4 to 72% for this experiment.



(a) GNB confusion matrix

(b) E1 DT confusion matrix

(c) E1 SVC confusion matrix

(d) DNN confusion matrix

Figure 4.10: E5 Confusion matrices for E1 models (N = 171).

	Precision	Recall	F-1 Score	N
Not removed	0.27	0.70	0.39	37
Removed	0.86	0.49	0.62	134
Micro avg	0.53	0.53	0.53	171
Macro avg	0.56	0.59	0.51	171
Weighted avg	0.73	0.53	0.57	171

Table 4.37: E5 GNB Classification result, using E1 model (N=171).

	Precision	Recall	F-1 Score	N
Not removed	0.26	0.76	0.39	37
Removed	0.86	0.40	0.55	134
Micro avg	0.48	0.48	0.48	171
Macro avg	0.56	0.58	0.47	171
Weighted avg	0.73	0.48	0.51	171

Table 4.38: E5 DT Classification result, using E1 weighted model (N = 171).

	Precision	Recall	F-1 Score	N
Not removed	0.39	0.51	0.44	37
Removed	0.85	0.78	0.81	134
Micro avg	0.72	0.72	0.72	171
Macro avg	0.62	0.64	0.63	171
Weighted avg	0.75	0.72	0.73	171

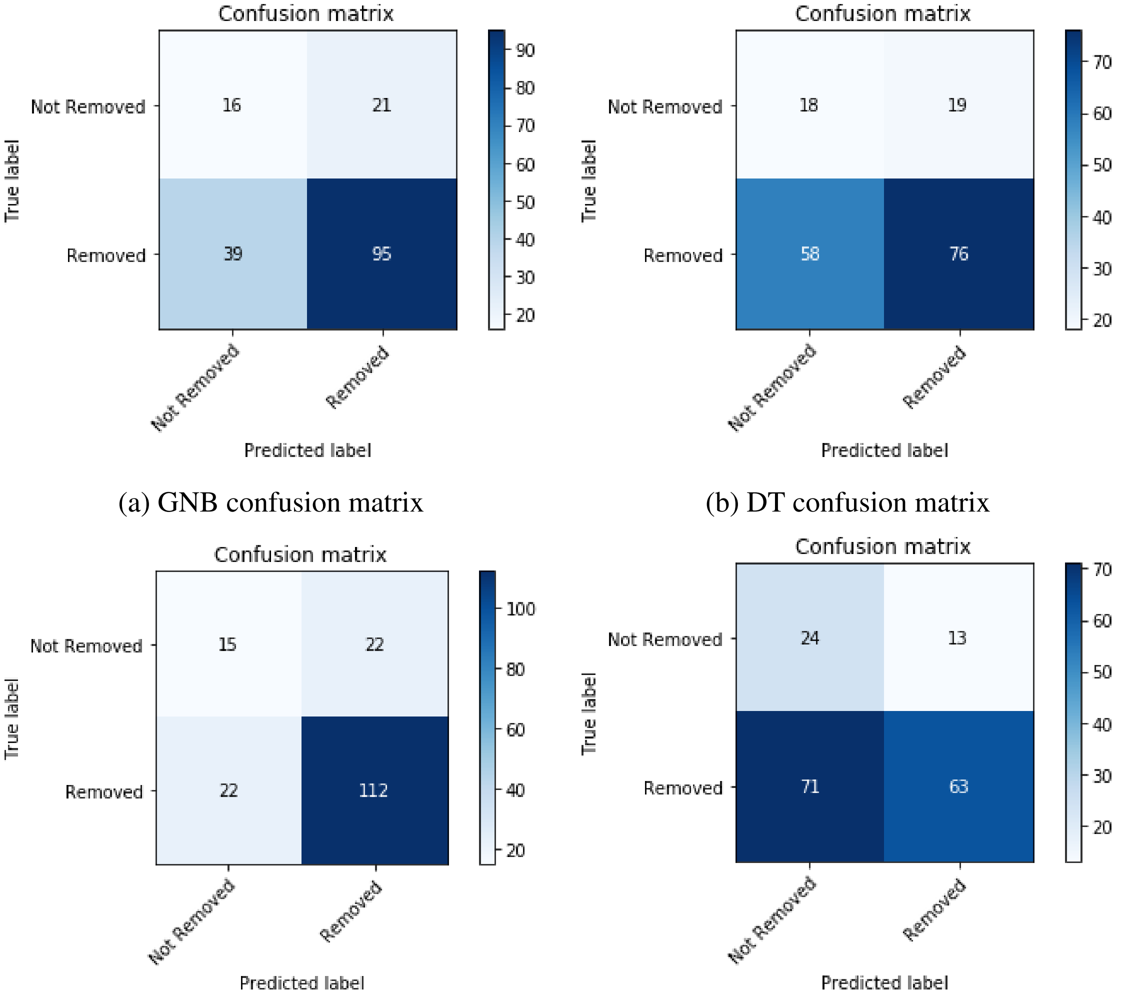
Table 4.39: E5 SVC Classification result, using E1 weighted model (N = 171).

	Precision	Recall	F-1 Score	N
Not removed	0.23	0.86	0.36	37
Removed	0.84	0.20	0.33	134
Micro avg	0.35	0.35	0.35	171
Macro avg	0.54	0.53	0.34	171
Weighted avg	0.71	0.35	0.33	171

Table 4.40: E5 DNN Classification result, using E1 model (N = 171).

4.8.2 E2 Model Performance

The SVC had the highest accuracy of the models from E2, likely due to its tendency to label comments as Removed. As shown in Figure 4.4c, with the E2 validation set this tendency created many false positives, but as E5 is majority removed, this led to higher accuracy. The remaining models tended toward false negatives for the Removed class, in particular the DNN and DT, which can be seen in Figure 4.11. All of the E2 models performed worse on the E5 set compared to E4, which suggests that although the models were not trained on the annotator labels, the annotator decisions are easier to predict than the original tag.



(a) GNB confusion matrix

(b) DT confusion matrix

(c) SVC confusion matrix

(d) DNN confusion matrix

Figure 4.11: E5 confusion matrices for E2 models (N = 171).

	Precision	Recall	F-1 Score	N
Not removed	0.29	0.43	0.35	37
Removed	0.82	0.71	0.76	134
Micro avg	0.65	0.65	0.65	171
Macro avg	0.55	0.57	0.55	171
Weighted avg	0.70	0.65	0.67	171

Table 4.41: E5 GNB Classification results, using E2 model (N = 171).

	Precision	Recall	F-1 Score	N
Not removed	0.24	0.49	0.32	37
Removed	0.80	0.57	0.66	134
Micro avg	0.55	0.55	0.55	171
Macro avg	0.52	0.53	0.49	171
Weighted avg	0.68	0.55	0.59	171

Table 4.42: E5 DT Classification results, using E2 model (N = 171).

	Precision	Recall	F-1 Score	N
Not removed	0.41	0.41	0.41	37
Removed	0.84	0.84	0.84	134
Micro avg	0.74	0.74	0.74	171
Macro avg	0.62	0.62	0.62	171
Weighted avg	0.74	0.74	0.74	171

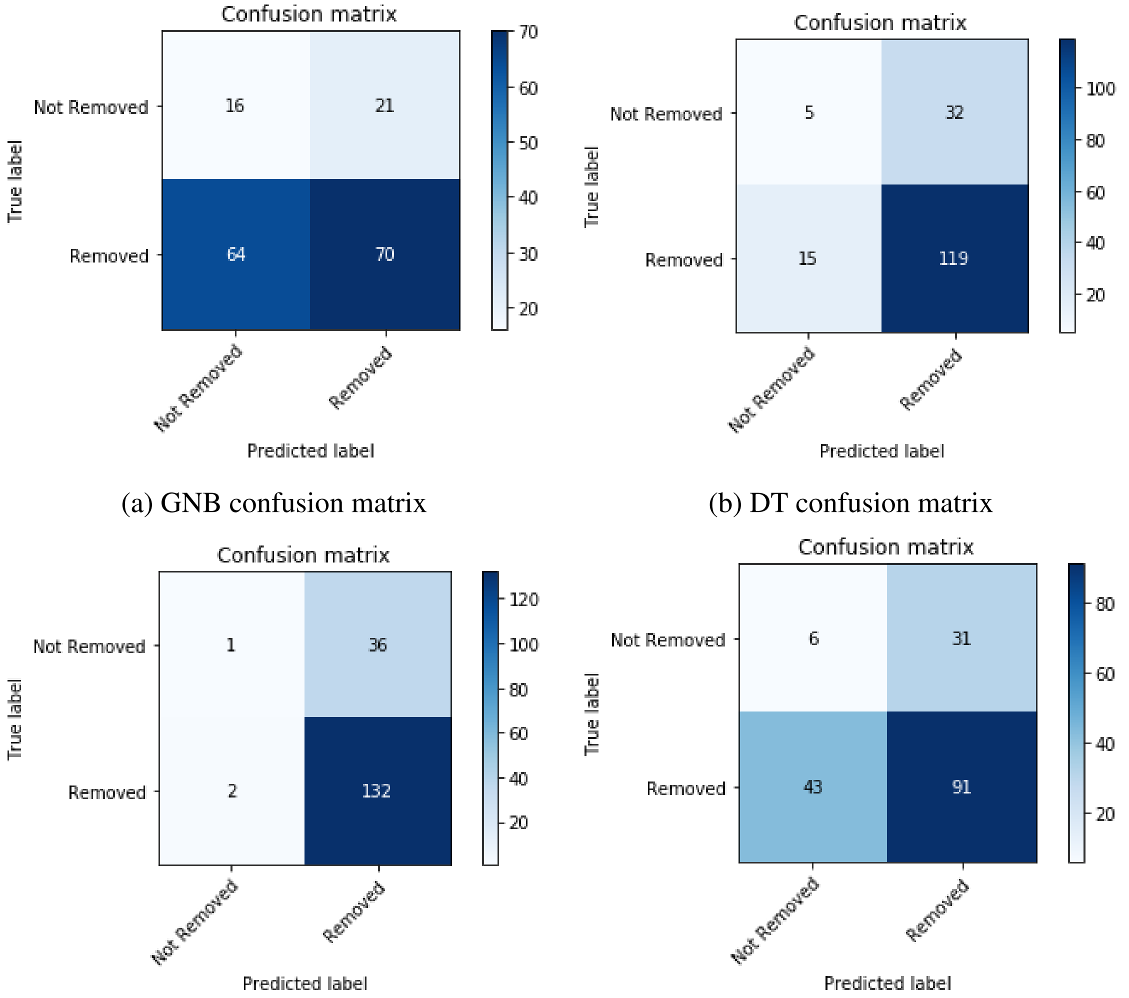
Table 4.43: E5 SVC Classification result, using E2 model (N = 171).

	Precision	Recall	F-1 Score	N
Not removed	0.25	0.65	0.36	37
Removed	0.83	0.47	0.60	134
Micro avg	0.51	0.51	0.51	171
Macro avg	0.54	0.56	0.48	171
Weighted avg	0.70	0.51	0.55	171

Table 4.44: E5 DNN Classification result, using E2 model (N = 171).

4.8.3 E3 Model Performance

The DT and SVC models performed well on this subset at 73% and 78% overall accuracy. This is due to the models' tendency to label almost all comments as Removed, as seen in the confusion matrices in Figure 4.12. The DT and SVC therefore have a high false positive rate, labeling many Not Removed comments as Removed. The DNN and GNB, at a lower accuracy of 57% and 50% each, had a higher tendency to create false negatives. This can be seen within the GNB and DT models by comparing the recall scores for the Not Removed class in Tables 4.45 and 4.46.



(a) GNB confusion matrix

(b) DT confusion matrix

(c) SVC confusion matrix

(d) DNN confusion matrix

Figure 4.12: E5 confusion matrices for E3 models (N = 171).

	Precision	Recall	F-1 Score	N
Not removed	0.20	0.43	0.27	37
Removed	0.77	0.52	0.62	134
Micro avg	0.50	0.50	0.50	171
Macro avg	0.48	0.48	0.45	171
Weighted avg	0.65	0.50	0.55	171

Table 4.45: E5 GNB Classification result, using E3 model (N = 171).

	Precision	Recall	F-1 Score	N
Not removed	0.25	0.14	0.18	37
Removed	0.79	0.89	0.84	134
Micro avg	0.73	0.73	0.73	171
Macro avg	0.52	0.51	0.51	171
Weighted avg	0.67	0.73	0.69	171

Table 4.46: E5 DT Classification result, using E3 weighted model (N = 171).

	Precision	Recall	F-1 Score	N
Not removed	0.33	0.03	0.05	37
Removed	0.79	0.99	0.87	134
Micro avg	0.78	0.78	0.78	171
Macro avg	0.56	0.51	0.46	171
Weighted avg	0.69	0.78	0.70	171

Table 4.47: E5 SVC Classification result, using E3 weighted model (N = 171).

	Precision	Recall	F-1 Score	N
Not removed	0.12	0.16	0.14	37
Removed	0.75	0.68	0.71	134
Micro avg	0.57	0.57	0.57	171
Macro avg	0.43	0.42	0.43	171
Weighted avg	0.61	0.57	0.59	171

Table 4.48: E5 DNN Classification result, using E3 model (N = 171).

4.9 E6 - Disagree (Original Tag)

This subset of comments was comprised of the comments for which annotators disagreed on a label, and is much smaller than the previous sets. The size of this subset was 51 comments, and it had a distribution of 76.5% Removed and 23.5% Not Removed.

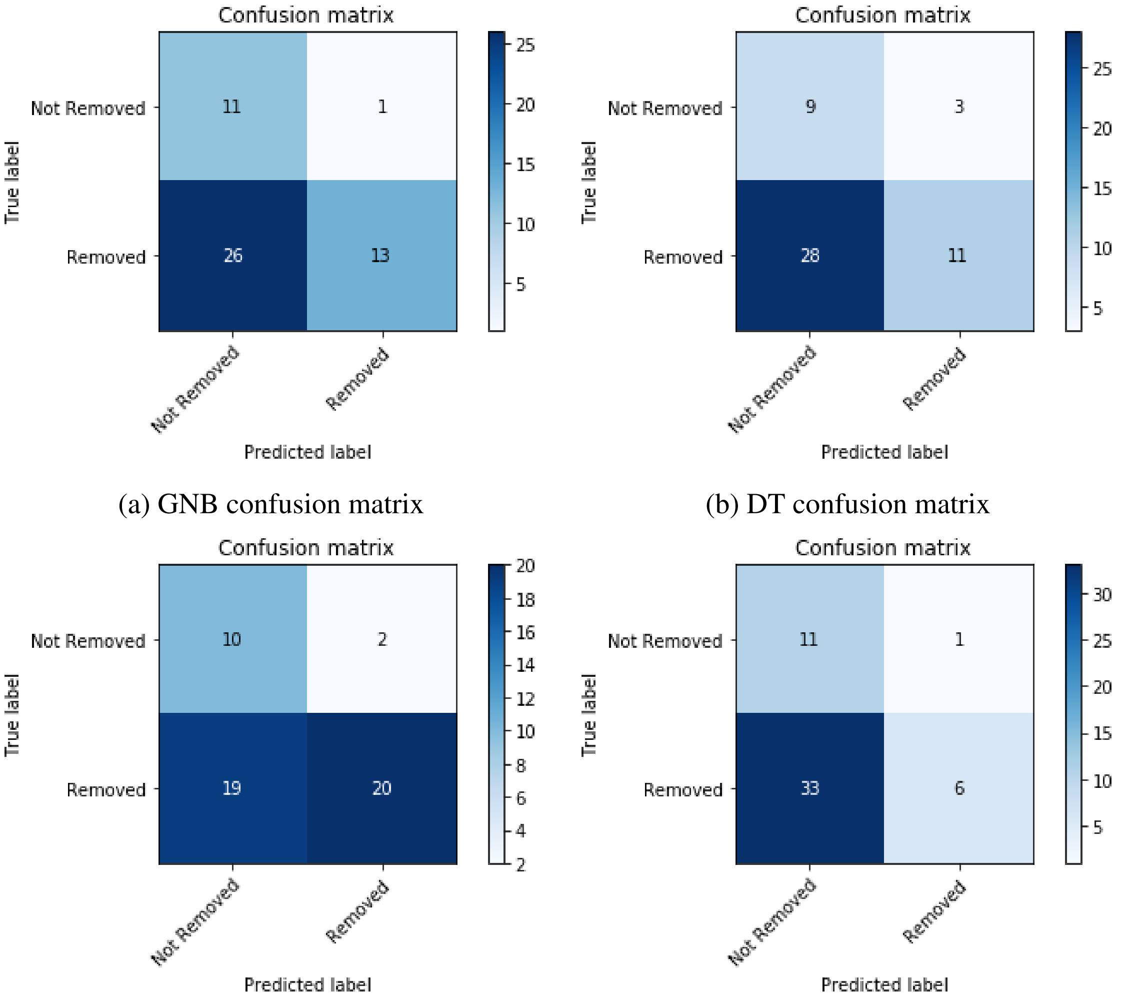
	GNB	DT	SVC	DNN
E1	47	39	**59**	33
E2	**75**	53	65	51
E3	49	65	**78**	63

Table 4.49: Percent accuracy for E1, E2, and E3 models across E6 set.

The SVC continued to outperform the other models for each experiment; however, there was a visible drop in performance for all E1 models compared to E4 and E5. Surprisingly, the E2 GNB outperformed all other E2 models, a key difference from the SVC models' consistently higher accuracy for E4-E6 otherwise.

4.9.1 E1 Model Performance

Model accuracy across this set lessened further for all models, suggesting that these comments, which were difficult for humans, were also difficult for the models. Additionally, all of the models presented almost exclusively false negatives for the Removed class for their errors. This suggests that for that for difficult classifications such as this subset, the models would choose Not Removed for the class.



(a) GNB confusion matrix

(b) DT confusion matrix

(c) SV confusion matrix

(d) DNN confusion matrix

Figure 4.13: E6 Confusion matrices for E1 models (N = 51).

	Precision	Recall	F-1 Score	N
Not removed	0.30	0.92	0.45	12
Removed	0.93	0.33	0.49	39
Micro avg	0.47	0.47	0.47	51
Macro avg	0.61	0.62	0.47	51
Weighted avg	0.78	0.47	0.48	51

Table 4.50: E6 GNB Classification result, using E1 model (N = 51).

	Precision	Recall	F-1 Score	N
Not removed	0.24	0.75	0.37	12
Removed	0.79	0.28	0.42	39
Micro avg	0.39	0.39	0.39	51
Macro avg	0.51	0.52	0.39	51
Weighted avg	0.66	0.39	0.40	51

Table 4.51: E6 DT Classification result, using weighted E1 model (N = 51).

	Precision	Recall	F-1 Score	N
Not removed	0.34	0.83	0.49	12
Removed	0.91	0.51	0.66	39
Micro avg	0.59	0.59	0.59	51
Macro avg	0.63	0.67	0.57	51
Weighted avg	0.78	0.59	0.62	51

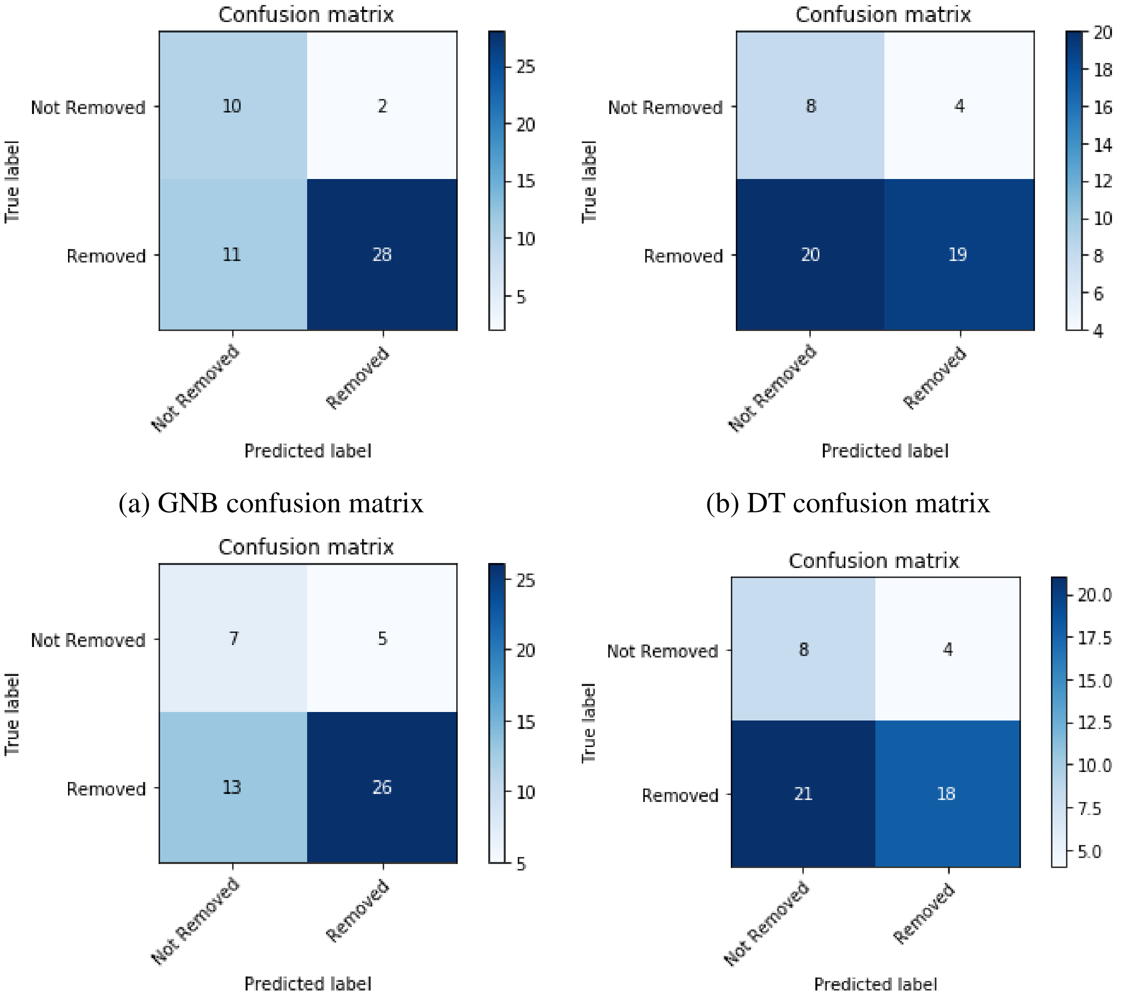
Table 4.52: E6 SVC Classification results, using weighted E1 model (N = 51).

	Precision	Recall	F-1 Score	N
Not removed	0.25	0.92	0.39	12
Removed	0.86	0.15	0.26	39
Micro avg	0.33	0.33	0.33	51
Macro avg	0.55	0.54	0.33	51
Weighted avg	0.71	0.33	0.29	51

Table 4.53: E6 DNN Classification results, using weighted E1 model (N = 51).

4.9.2 E2 Model Performance

The GNB and SVC models performed much better at 75% and 65% accuracy, respectively, than the remaining models. Both the DNN and DT models experienced a large amount of false negatives for the Removed class, as shown in Figure 4.14. The accuracy of the DT and DNN were lower, at 53% and 51% respectively.



(a) GNB confusion matrix

(b) DT confusion matrix

(c) SVC confusion matrix

(d) DNN confusion matrix

Figure 4.14: E6 confusion matrices for E2 models (N = 51).

	Precision	Recall	F-1 Score	N
Not removed	0.48	0.83	0.61	12
Removed	0.93	0.72	0.81	39
Micro avg	0.75	0.75	0.75	51
Macro avg	0.70	0.78	0.71	51
Weighted avg	0.83	0.75	0.76	51

Table 4.54: E6 GNB Classification results, using E2 model (N = 51).

	Precision	Recall	F-1 Score	N
Not removed	0.29	0.67	0.40	12
Removed	0.83	0.49	0.61	39
Micro avg	0.53	0.53	0.53	51
Macro avg	0.56	0.58	0.51	51
Weighted avg	0.70	0.53	0.56	51

Table 4.55: E6 DT Classification results, using E2 model (N = 51).

	Precision	Recall	F-1 Score	N
Not removed	0.35	0.58	0.44	12
Removed	0.84	0.67	0.74	39
Micro avg	0.65	0.65	0.65	51
Macro avg	0.59	0.62	0.59	51
Weighted avg	0.72	0.65	0.67	51

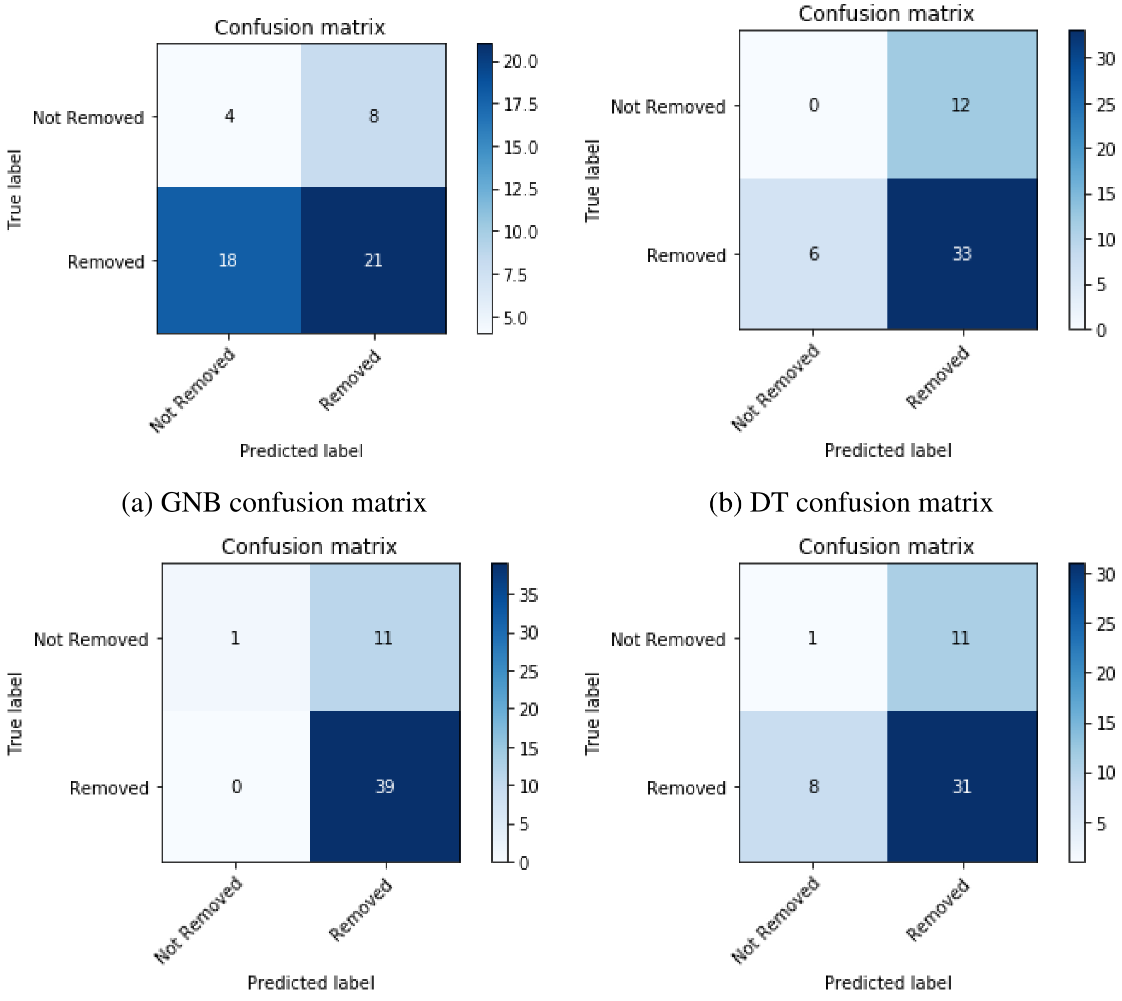
Table 4.56: E6 SVC Classification results, using E2 model (N = 51).

	Precision	Recall	F-1 Score	N
Not removed	0.28	0.67	0.39	12
Removed	0.82	0.46	0.59	39
Micro avg	0.51	0.51	0.51	51
Macro avg	0.55	0.56	0.49	51
Weighted avg	0.69	0.51	0.54	51

Table 4.57: E6 DNN Classification result, using E2 model (N = 51).

4.9.3 E3 Model Performance

The accuracy of these models is high in part due to the nature of the E3 models; many of them almost exclusively select the majority class for the dataset, which in this case is Removed. As this is also the majority of the E6 subset, these models perform well for the overall accuracy but fall short for other metrics. The confusion matrices in Figure 4.15 show that while all of these E3 models are able to correctly identify most of the Removed class, there are many false positives for the Removed class from the DT, SVC, and DNN models. The GNB model, also at a much lower accuracy at 49%, had many false negatives.



(a) GNB confusion matrix

(b) DT confusion matrix

(c) SVC confusion matrix

(d) DNN confusion matrix

Figure 4.15: E6 confusion matrices for E3 models (N = 51).

	Precision	Recall	F-1 Score	N
Not removed	0.18	0.33	0.24	12
Removed	0.72	0.54	0.62	39
Micro avg	0.49	0.49	0.49	51
Macro avg	0.45	0.44	0.43	51
Weighted avg	0.60	0.49	0.53	51

Table 4.58: E6 GNB Classification result, using E3 model (N = 51).

	Precision	Recall	F-1 Score	N
Not removed	0.00	0.00	0.00	12
Removed	0.73	0.85	0.79	39
Micro avg	0.65	0.65	0.65	51
Macro avg	0.37	0.42	0.39	51
Weighted avg	0.56	0.65	0.60	51

Table 4.59: E6 DT Classification result, using E3 weighted model (N = 51).

	Precision	Recall	F-1 Score	N
Not removed	1.00	0.08	0.15	12
Removed	0.78	1.00	0.88	39
Micro avg	0.78	0.78	0.78	51
Macro avg	0.89	0.54	0.52	51
Weighted avg	0.83	0.78	0.71	51

Table 4.60: E6 SVC Classification result, using E3 weighted model (N = 51).

	Precision	Recall	F-1 Score	N
Not removed	0.11	0.08	0.10	39
Removed	0.74	0.79	0.77	12
Micro avg	0.63	0.63	0.63	51
Macro avg	0.42	0.44	0.43	51
Weighted avg	0.59	0.63	0.61	51

Table 4.61: E6 DNN Classification result, using E3 model (N = 51).

4.10 Interpretation of Model Behaviors using Moderation Study Results

The creation of a novel dataset of expertly twice-annotated comments allowed for a unique measurement of the performance of the created ML models. The large variation of performance across these subsets highlighted the potential flaws in models that rely on the raw, original labels from the dataset of Reddit comments as there are potentially many mislabeled comments within the original labels. As these original labels used in the training sets often included comments that were never seen by a human moderator as well as comments that were *nuked*, there were many false positives and negatives that may have influenced the training of the model. In this context, *nuking* refers to the process of removing an entire comment chain at once, which often removes otherwise acceptable comments within the chain.

The results of the ML models reinforce that classifying text, even for a community like */r/science* that has strict rules regarding conduct, is a difficult task. Survey respondents varied greatly in their moderation style and the types of content that they prefer to moderate, showing subjectivity to the task among these expert annotators; most annotators were of a higher rank than average and spent several years moderating. Within the subset of twice-annotated comments, they agreed with each other only 77% of the time, highlighting a large variation present even within a community of experts that have been moderating for several years.

Since moderators often use context to determine whether a comment should be removed, it was a serious limitation of the project that context was not available to participants - or to the machine learning models. The number of comments on a popular */r/science* post can easily reach several thousand, and the amount of these comments that exist as a top-level comment with no context is relatively small. Respondents for the survey often mentioned context within their responses, as shown in Section 3.3, especially for comments in which annotators disagreed. The inclusion of context for both the annotators and the models could further improve the accuracy of the models and agreement within annotators, but at the cost of additional time and complexity for the machine learning models.

4.11 Summary

Six separate experiments were conducted, three of which trained and evaluated models over varying distributions of data, and three that leveraged annotation data to further evaluate the trained models. As shown in Table 4.3, the SVC classifier outperformed all other models in almost all situations, although the E3 SVC and DT models suffered from the bias of the dataset and selected the majority class almost exclusively. The survey and annotation results suggest that moderating */r/science* content is a subjective task, hard for humans, and is reflected in the difficulty that the classification models faced.

Chapter 5

Conclusions

Multiple machine learning models were created to distinguish acceptable from unacceptable comments within the original corpus of random */r/science* comments. Survey data from participants was used to understand how moderators view the problem, and showed the high variation between experts as to how they approach the task of moderating */r/science*. This variation directly affected the created models, as the performance over the E6 (Disagree) set was notably lower than E4 and E5.

Due to constraints and the availability of participants, the amount of annotated data for the moderation task was restrictively small. With more incentives for participation, such as compensation, additional annotations could be collected to create a large pool of expertly annotated data to further the performance of the machine learning models. With such limited data to be used as the gold standard, the overall ability of the model was restricted by the reliability of the original labels, which could include thousands of comments that were never seen and judged by a moderator. Further training the models over expert annotations could help fix these issues, and weighting these samples heavier than the original labels could mitigate the imbalance from having a small subset of annotated comments. Additionally, given a larger sample of annotation explanations and survey responses, topic modeling can be performed to further understand moderation methods [8].

With additional training and data, the study and resulting models could be used to augment the AutoMod system that is currently in use by moderators. As AutoMod relies on a keyword-based regex system, the created models were likely more flexible and can identify comments that circumvent the strict set of keywords that are recognized by AutoMod.

Within the eight thousand comments used in this study, AutoMod accounted for nearly 30% of all Removed comments. While many offending comments would be too nuanced for AutoMod, further investigation of the created models could reveal additional keywords to be added to AutoMod. Furthermore, the longterm addition of a flagging system based on the models used for this study could enhance the ability of AutoMod and ensure that fewer comments fall through the cracks. By additionally adjusting the created models for online learning, a moderator team could pre-train a model on all past removed comments and a select subset of approved comments, and continually update and train the model as new comments are removed.

Moreover, efforts to visualize the classification models could illuminate the inner workings of these models and further understanding of the decisions they make. Attempts to understand and visualize these models are often limited by the domain of the problem or data type from which they originated. Visual data can provide intuition for observers to translate the model visualization to the data in question. Several researchers have established a relevance metric that can be used to quantify the importance of a unit (e.g. a word or pixel) of input to the final decision at the output layer [3, 34]. The ability to visualize an arbitrarily deep level of the network was first proposed by Erhan et al. in 2009, and is flexible enough to handle multiple architectures [16]. Yosinki et al. (2015) later provide two novel tools for visualization, but they are limited to convolutional neural networks [41]. These tools provide an extremely detailed view of their chosen model; one has seven different ways of visualizing the data on screen for the user at a given time, and the other provides a simplified view of the activation areas for the original raw image when run through a convolutional neural network. Visualization of Neural Machine Translation was recently explored by Ding et al. (2017), by way of layer-wise relevance propagation [15]. The authors create a relevance metric for use in a grayscale visualization of the relevance each word has to others, both in the original source sentence and in the preceding target words. This also allows the researcher to better understand why translation errors occur, as the relevance of the other words involved can be seen at a glance.

With additional annotated data, the models can also be further developed to identify prominent reasons for removal. Model performance and key defining features could be visualized to understand what lexicons or semantic structures can be used to differentiate passable comments from removals. Feedback from the surveys can be compared with the created models to evaluate how well the models capture human moderation behaviors for this extension. In the future, annotators could potentially opt-in to an internet browser extension that allows them to provide a removal reason directly to the database when they perform their normal moderation activities on Reddit. This could alleviate the large time commitment from the annotation task as well as allow for the use of context within decision-making, potentially improving the quality of the annotations. The ability for an automated moderation system to detect the reason for removal could additionally allow for this feedback to be passed on to the user whose comment was removed immediately, reducing the amount of messages the moderators receive asking for feedback.

Bibliography

[1] Reddit comment data. https://bigquery.cloud.google.com/dataset/fh-bigquery:reddit_comments. Accessed 15 November, 2017.

[2] Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 183–194, New York, NY, USA, 2008. ACM.

[3] Sebastian Bach, Alexander Binder, Grgoire Montavon, Frederick Klauschen, Klaus-Robert Mller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10:1–46, 07 2015.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[5] David Banks. Aspects of the development of grammatical metaphor in scientific writing. *Cahiers de l'APLIUT*, 19(1):5–25, 1999.

[6] Chad Birch. What is automoderator?, 2012. https://www.reddit.com/r/AutoModerator/comments/q11pu/what_is_automoderator/.

[7] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009.

[8] David M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[9] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 144–152, New York, NY, USA, 1992. ACM.

[10] Stevie Chancellor, Zhiyuan (Jerry) Lin, and Munmun De Choudhury. “This post will just get taken down”: Characterizing removed pro-eating disorder social media content. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI ’16, pages 1157–1162, New York, NY, USA, 2016. ACM.

[11] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. You can’t stay here: The efficacy of Reddit’s 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):31:1–31:22, December 2017.

[12] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. The bag of communities: Identifying abusive behavior online with preexisting internet data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3175–3187. ACM, 2017.

[13] François Chollet et al. Keras. `https://keras.io`, 2015.

[14] Denzil Correa and Ashish Sureka. Chaff from the wheat: Characterization and modeling of deleted questions on Stack Overflow. In *Proceedings of the 23rd international conference on World wide web*, pages 631–642. ACM, 2014.

[15] Yanzhuo Ding, Yang Liu, Huanbo Luan, and Maosong Sun. Visualizing and understanding neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1150–1159. ACL, 2017.

[16] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

[17] Martín Abadi et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[18] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

[19] Devin Gaffney and J Nathan Matias. Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus. *PloS one*, 13(7):e0200162, 2018.

[20] Hugh Gosden. Discourse functions of marked theme in scientific research articles. *English for Specific Purposes*, 11(3):207 – 224, 1992.

[21] Michael AK Halliday. On the language of physical science. *Registers of Written English: Situational Factors and Linguistic Features*, 5:162–178, 1988.

[22] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28*, pages 1693–1701. Curran Associates, Inc., 2015.

[23] Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. Improving word representations via global context and multiple word prototypes. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2012.

[24] Wang Ling, Yulia Tsvetkov, Silvio Amir, Ramon Fermandez, Chris Dyer, Alan W Black, Isabel Trancoso, and Chu-Cheng Lin. Not all contexts are created equal: Better word representations with variable attention. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1367–1372, 2015.

[25] J Nathan Matias. The civic labor of online moderators. In *Internet Politics and Policy conference, Oxford, United Kingdom*, IPP 2016, pages 1–10, Oxford, UK, 2016. Oxford Internet Insitute.

[26] J. Nathan Matias. Posting rules in online discussions prevents problems and increases participation, 2016. `https://civilservant.io/moderation_experiment_r_science_rule_posting.html`.

[27] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv e-prints*, January 2013.

[28] Richard Nate. Rhetoric in the early Royal Society. *Symbola et Emblemata*, 9:215–232, 1999.

[29] Casey Newton. The trauma floor: The secret lives of Facebook moderators in america, 2019. `https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona`.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[31] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[32] Max Pumperla. Hyperas. `https://github.com/maxpumperla/hyperas`, 2019.

[33] A. M. Rush, S. Chopra, and J. Weston. A neural attention model for abstractive sentence summarization. *arXiv e-prints*, September 2015.

[34] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv e-prints*, December 2013.

[35] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.

[37] W. Weaver. *Machine Translation of Languages*. MIT Press, Cambridge, MA, 1955.

[38] L. Weber and D. Seetharaman. The worst job in technology: Staring at human depravity to keep it off Facebook, 2017. `www.wsj.com/articles/the-worst-job-in-technology-staring-at-human-depravity-to-keep-it-off-facebook-1514398398`.

[39] Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.

[40] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of NAACL-HLT 2016*, NAACL-HLT '16, pages 1480–1489, Stroudsburg, PA, USA, 01 2016. Association for Computational Linguistics.

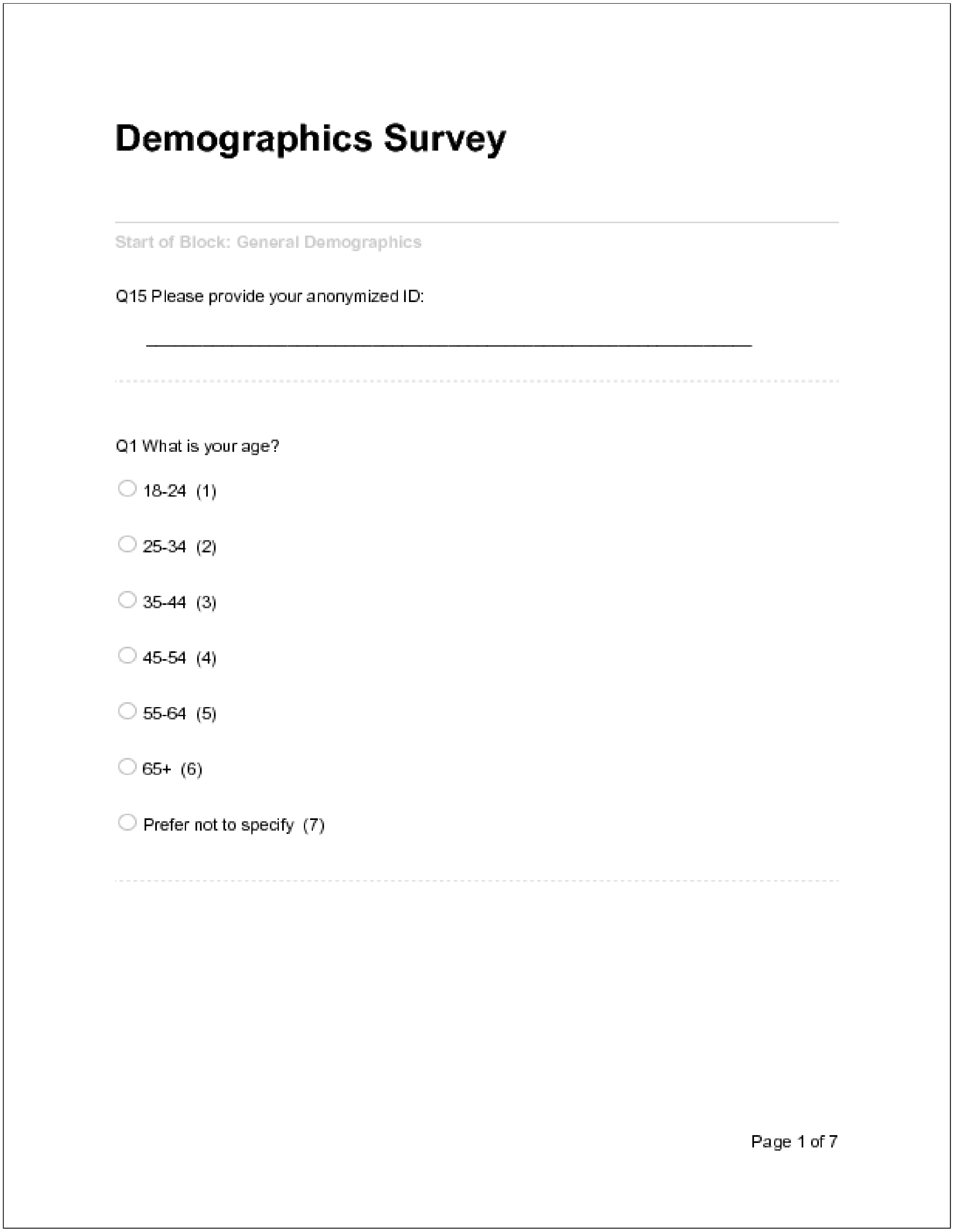
[41] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

Appendix A

Survey Instruments

This appendix includes the pages of the demographics survey that were distributed to all participants.

Figure A.1: Survey (page 1)



Demographics Survey

Start of Block: General Demographics

Q15 Please provide your anonymized ID:

____

Q1 What is your age?

○ 18-24 (1)

○ 25-34 (2)

○ 35-44 (3)

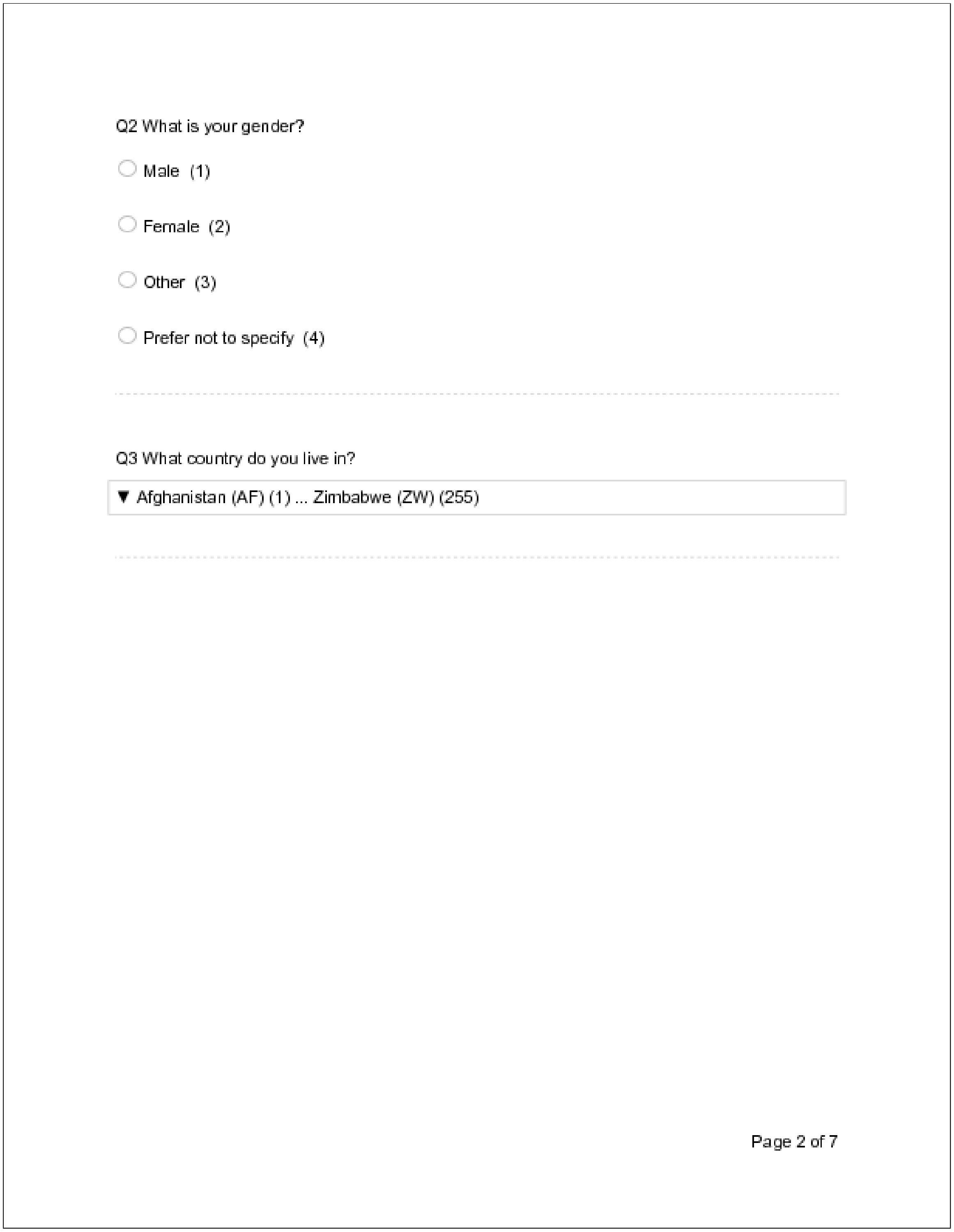
○ 45-54 (4)

○ 55-64 (5)

○ 65+ (6)

○ Prefer not to specify (7)

Figure A.2: Survey (page 2)



Q2 What is your gender?

○ Male (1)

○ Female (2)

○ Other (3)

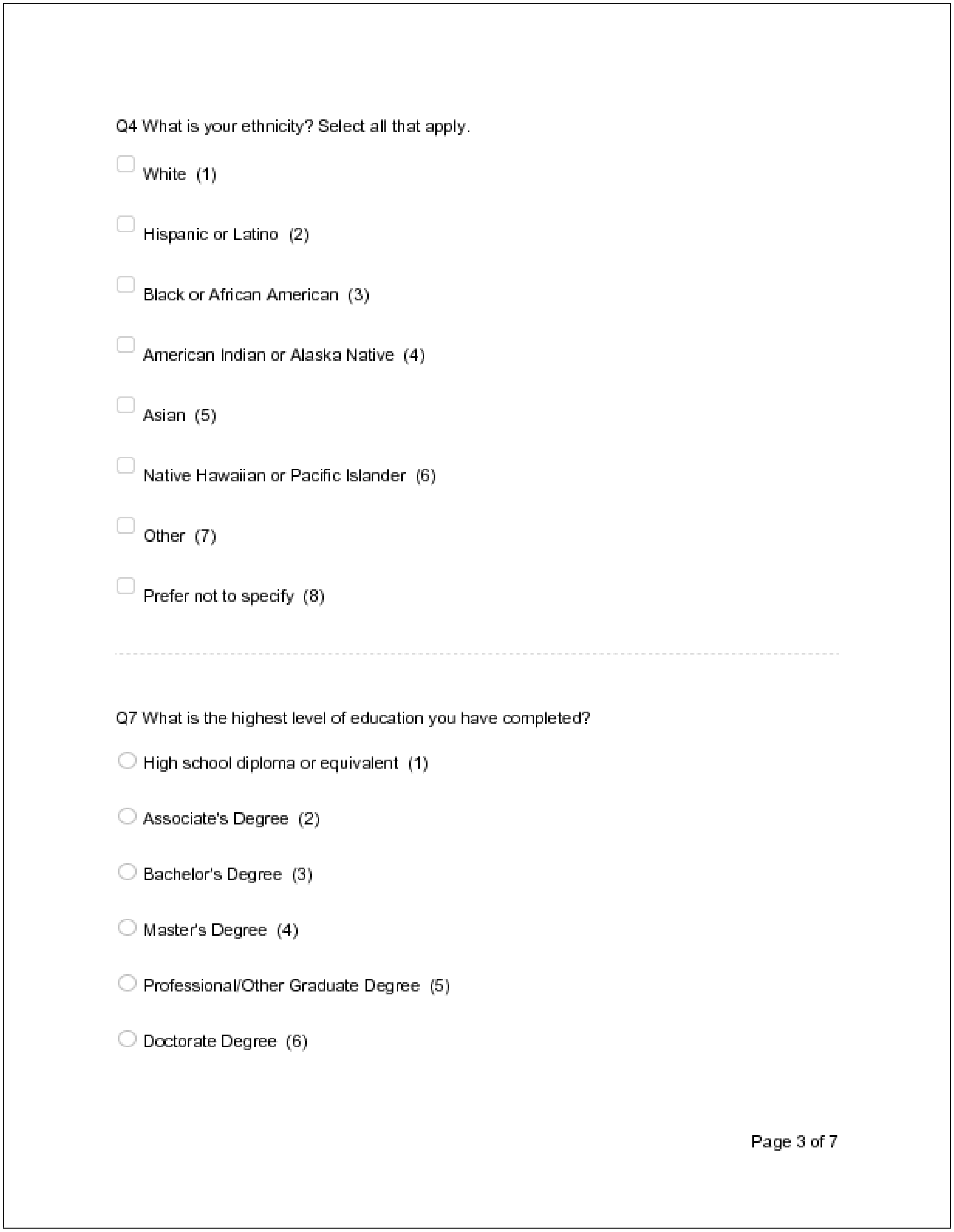
○ Prefer not to specify (4)

Q3 What country do you live in?

▼ Afghanistan (AF) (1) ... Zimbabwe (ZW) (255)

Page 2 of 7

Figure A.3: Survey (page 3)



Q4 What is your ethnicity? Select all that apply.

□ White (1)

□ Hispanic or Latino (2)

□ Black or African American (3)

□ American Indian or Alaska Native (4)

□ Asian (5)

□ Native Hawaiian or Pacific Islander (6)

□ Other (7)

□ Prefer not to specify (8)

Q7 What is the highest level of education you have completed?

○ High school diploma or equivalent (1)

○ Associate's Degree (2)

○ Bachelor's Degree (3)

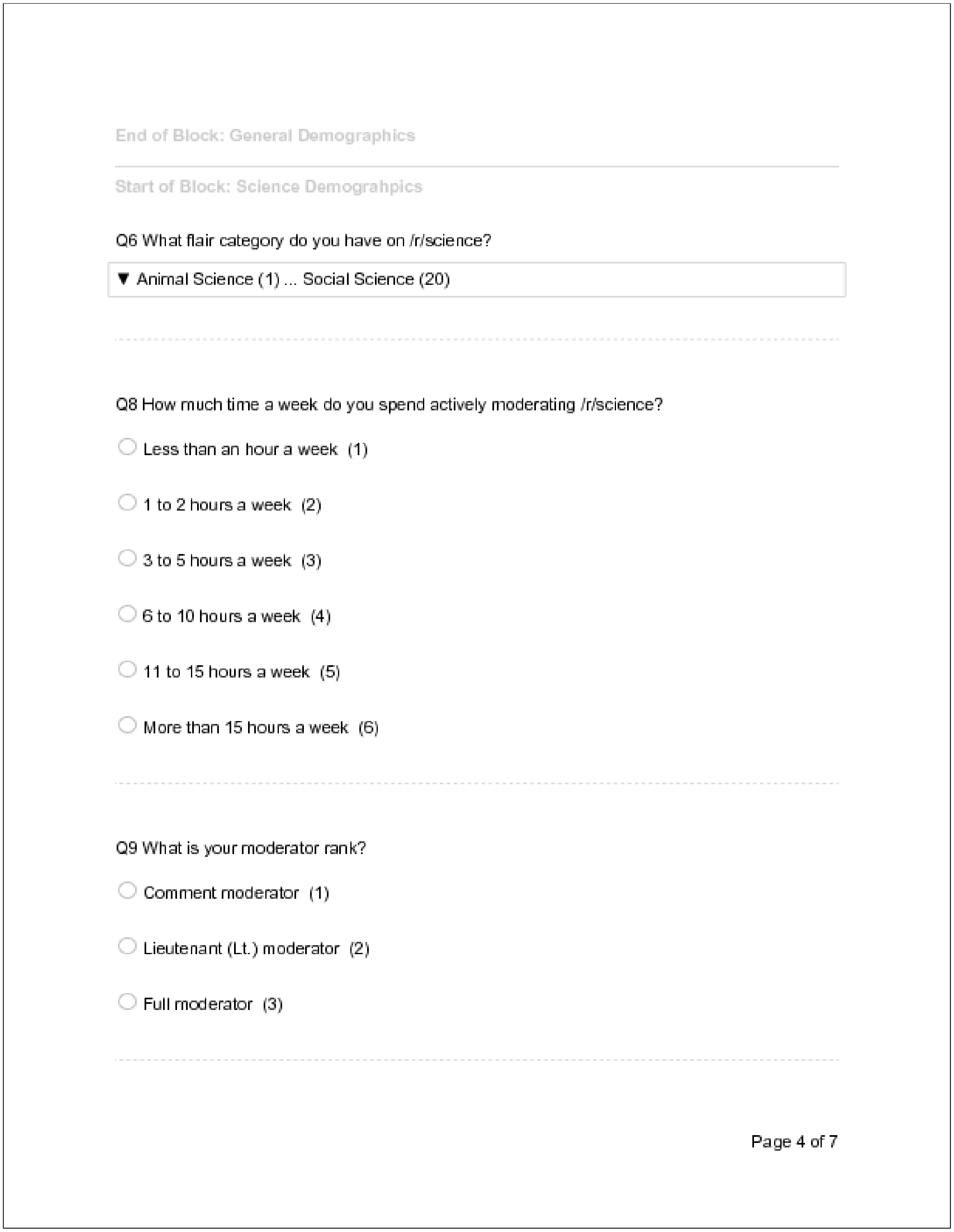
○ Master's Degree (4)

○ Professional/Other Graduate Degree (5)

○ Doctorate Degree (6)

Page 3 of 7

Figure A.4: Survey (page 4)



End of Block: General Demographics

Start of Block: Science Demograhpics

Q6 What flair category do you have on /r/science?

▼ Animal Science (1) ... Social Science (20)

Q8 How much time a week do you spend actively moderating /r/science?

○ Less than an hour a week (1)

○ 1 to 2 hours a week (2)

○ 3 to 5 hours a week (3)

○ 6 to 10 hours a week (4)

○ 11 to 15 hours a week (5)

○ More than 15 hours a week (6)

Q9 What is your moderator rank?

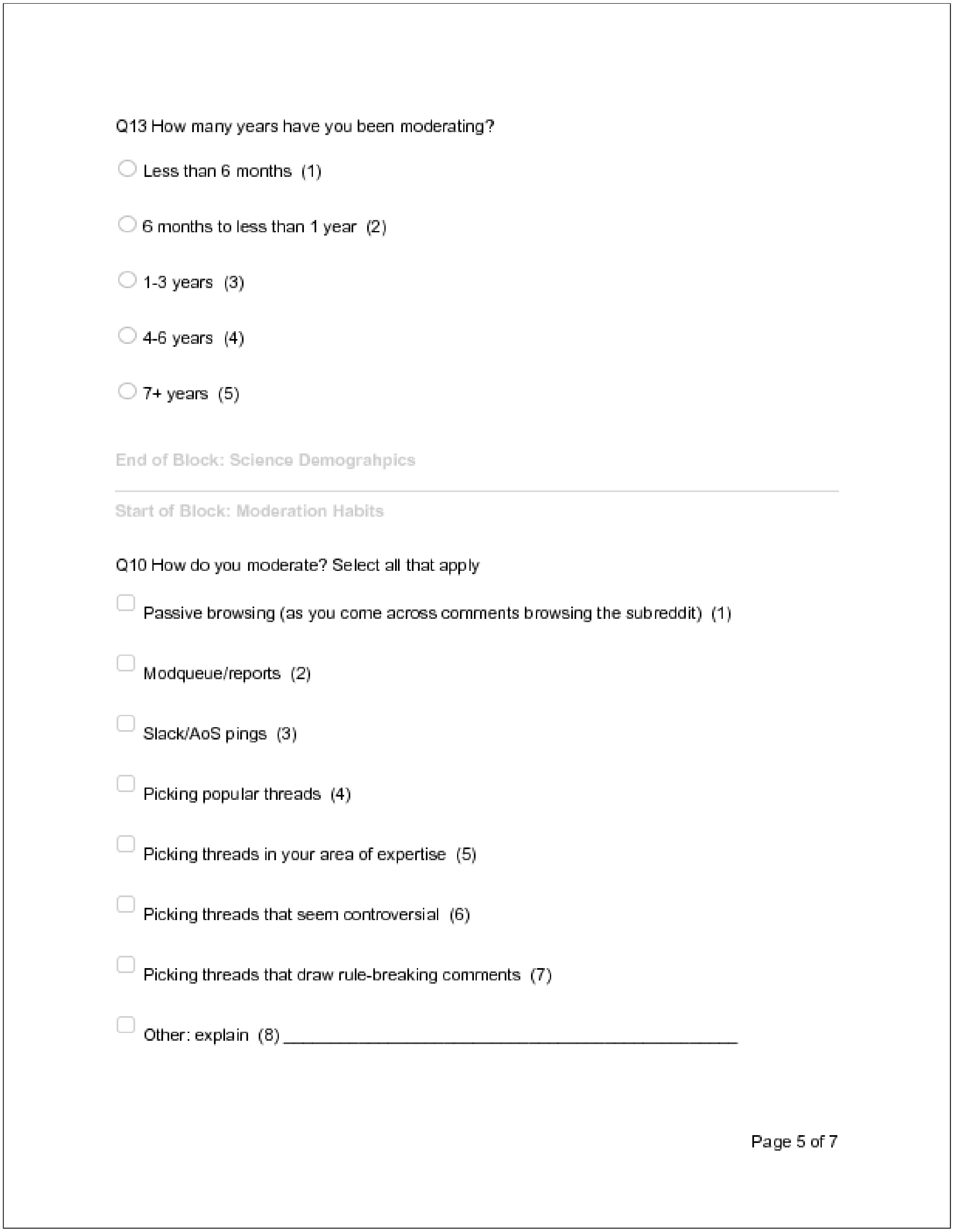
○ Comment moderator (1)

○ Lieutenant (Lt.) moderator (2)

○ Full moderator (3)

Page 4 of 7

Figure A.5: Survey (page 5)



Q13 How many years have you been moderating?

○ Less than 6 months (1)

○ 6 months to less than 1 year (2)

○ 1-3 years (3)

○ 4-6 years (4)

○ 7+ years (5)

End of Block: Science Demograhpics

Start of Block: Moderation Habits

Q10 How do you moderate? Select all that apply

□ Passive browsing (as you come across comments browsing the subreddit) (1)

□ Modqueue/reports (2)

□ Slack/AoS pings (3)

□ Picking popular threads (4)

□ Picking threads in your area of expertise (5)

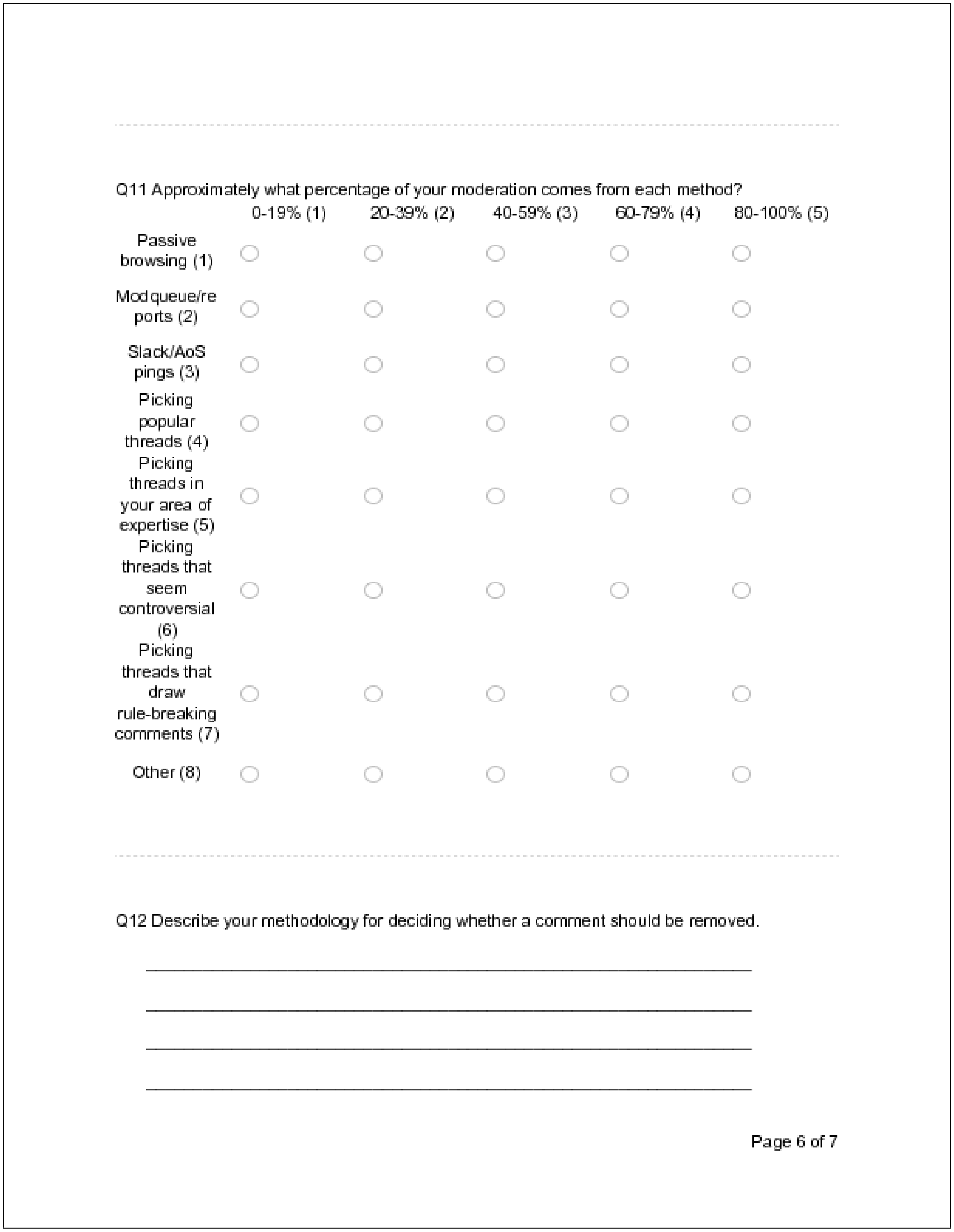
□ Picking threads that seem controversial (6)

□ Picking threads that draw rule-breaking comments (7)

□ Other: explain (8) ____

Page 5 of 7

Figure A.6: Survey (page 6)



Q11 Approximately what percentage of your moderation comes from each method?

	0-19% (1)	20-39% (2)	40-59% (3)	60-79% (4)	80-100% (5)
Passive browsing (1)	○	○	○	○	○
Modqueue/reports (2)	○	○	○	○	○
Slack/AoS pings (3)	○	○	○	○	○
Picking popular threads (4)	○	○	○	○	○
Picking threads in your area of expertise (5)	○	○	○	○	○
Picking threads that seem controversial (6)	○	○	○	○	○
Picking threads that draw rule-breaking comments (7)	○	○	○	○	○
Other (8)	○	○	○	○	○

Q12 Describe your methodology for deciding whether a comment should be removed.

____

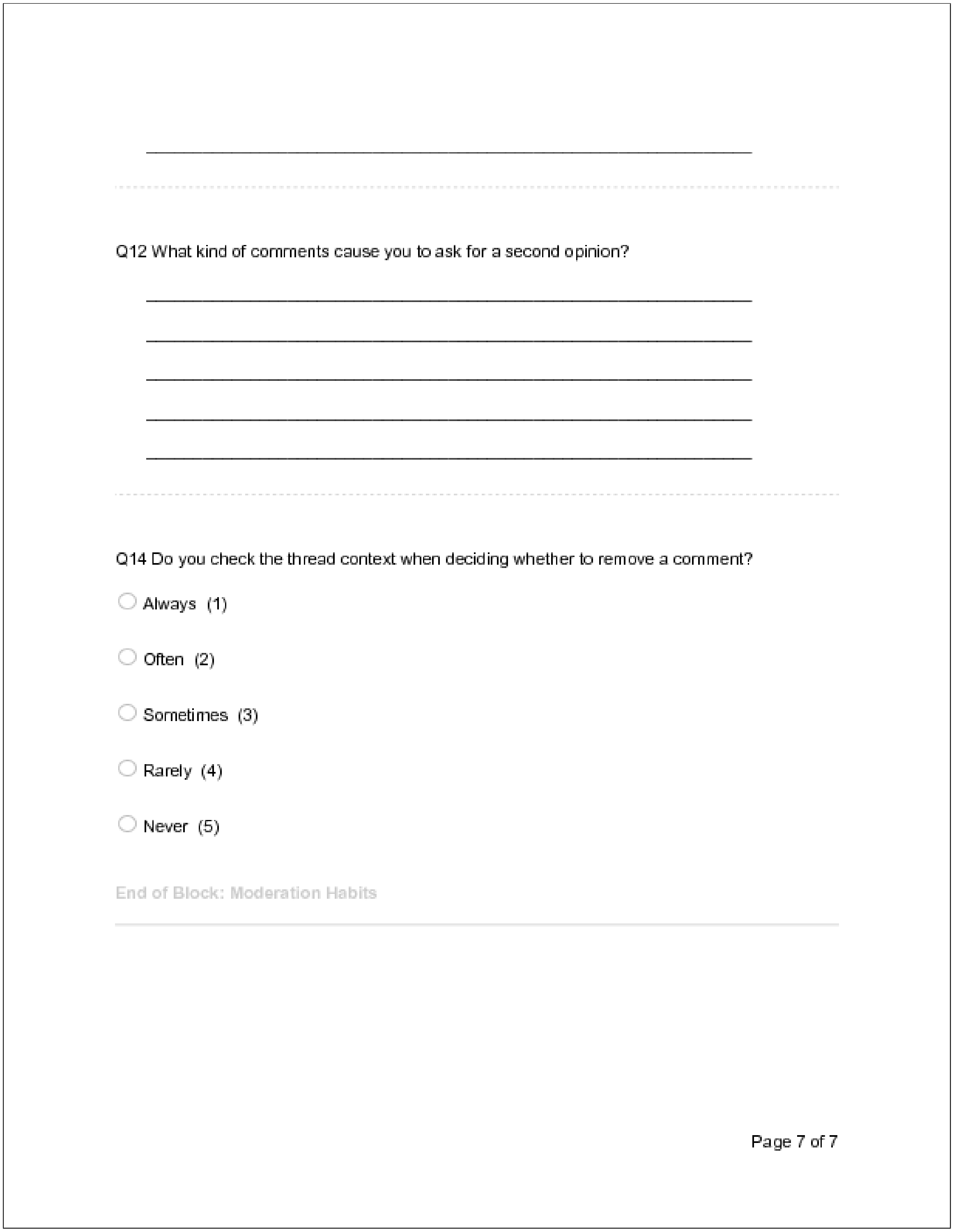
____

____

____

Page 6 of 7

Figure A.7: Survey (page 7)



____

Q12 What kind of comments cause you to ask for a second opinion?

____

____

____

____

____

Q14 Do you check the thread context when deciding whether to remove a comment?

○ Always (1)

○ Often (2)

○ Sometimes (3)

○ Rarely (4)

○ Never (5)

End of Block: Moderation Habits

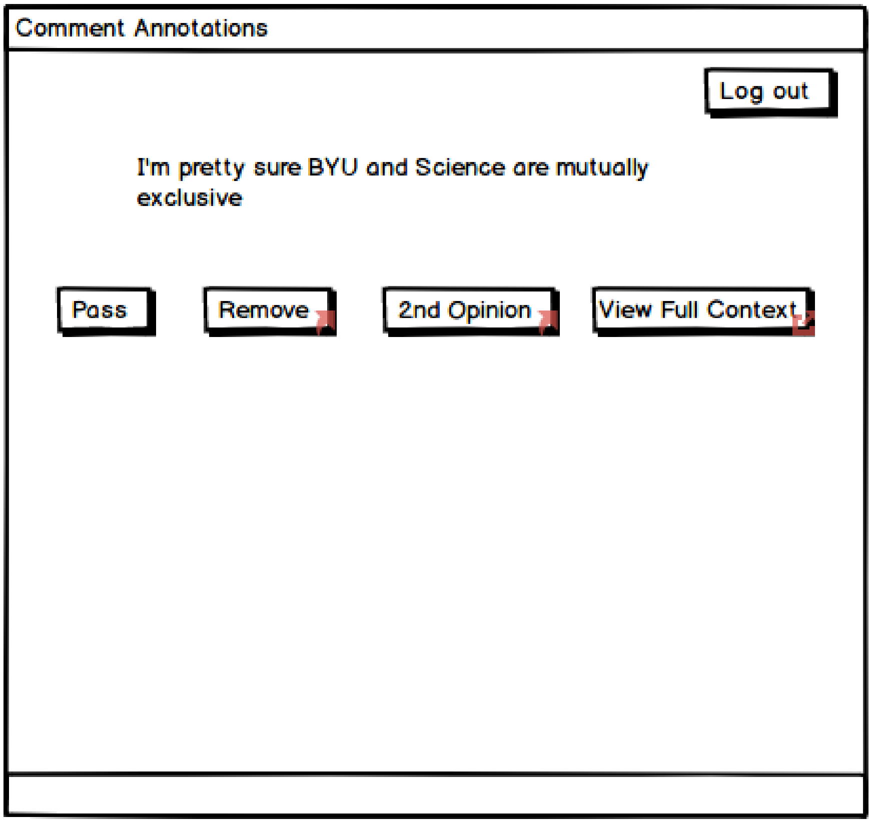
Page 7 of 7

Appendix B

Annotation Task Instruments

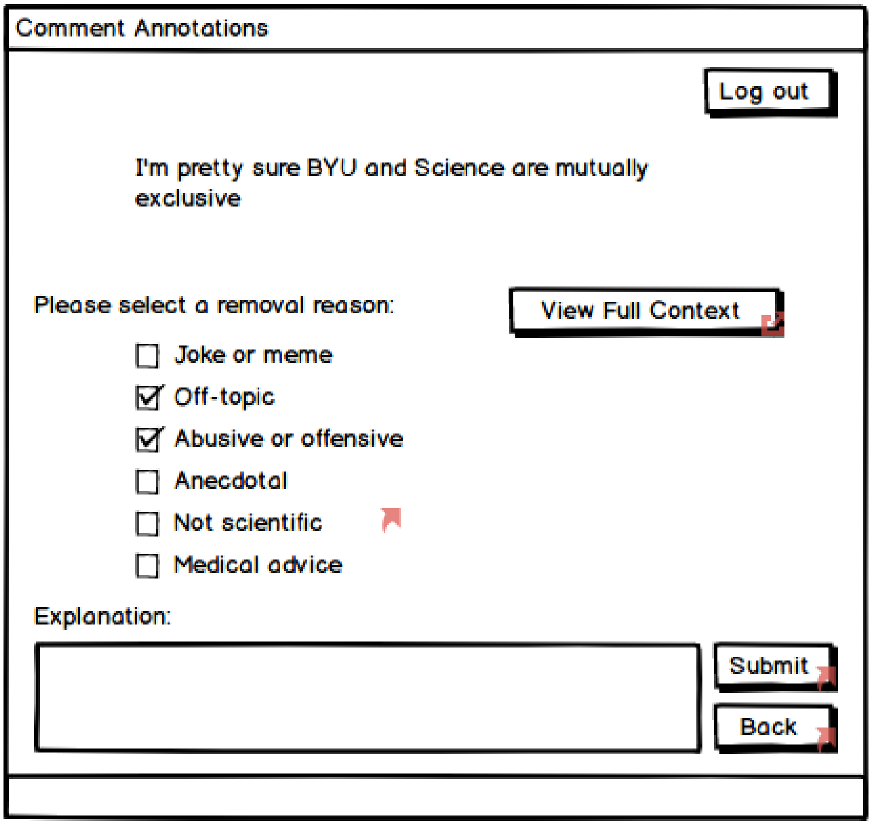
This appendix includes the draft and final versions of the annotation application. Wireframes are a visual representation of the application the participants used to record data, sketched before development, and are included as Figures B.1, B.2, and B.3. Screenshots from the live annotation task website are also included to demonstrate the interface the participants used, and are included as Figures B.4, B.5, and B.6.

Figure B.1: The main view of the application.



I'm pretty sure BYU and Science are mutually exclusive

Figure B.2: The user chooses the reasons for comment removal. The shown checkboxes for the removal reasons were not included in the final version, limiting the participants to one removal reason per comment.



Comment Annotations

Log out

I'm pretty sure BYU and Science are mutually exclusive

Please select a removal reason:

View Full Context

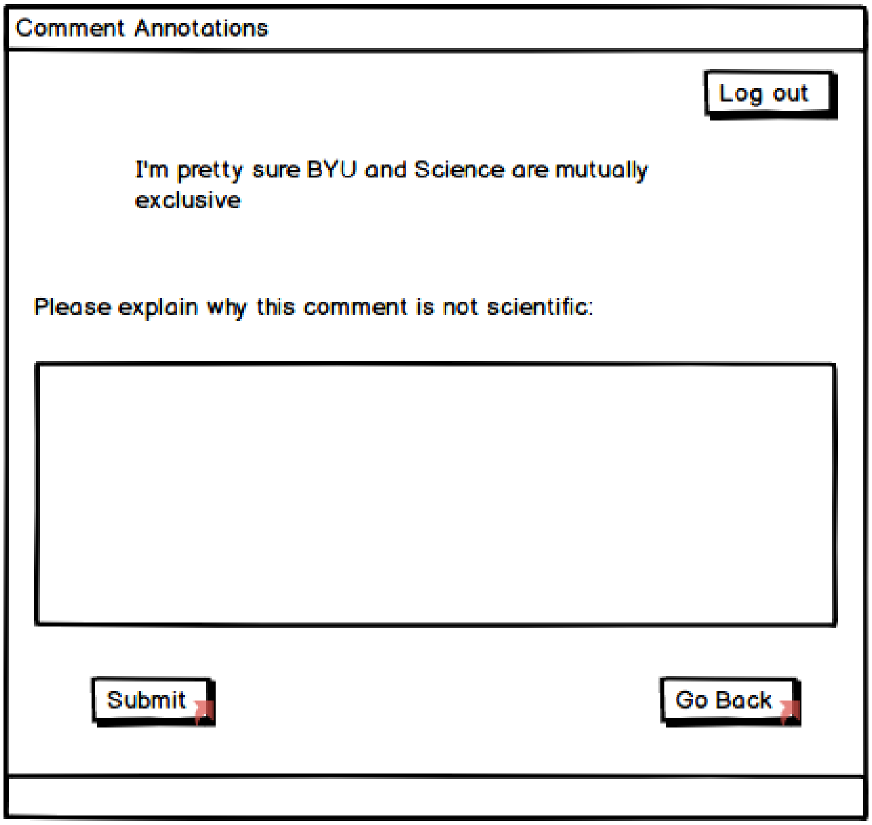
- □ Joke or meme
- ☑ Off-topic
- ☑ Abusive or offensive
- □ Anecdotal
- □ Not scientific
- □ Medical advice

Explanation:

Submit

Back

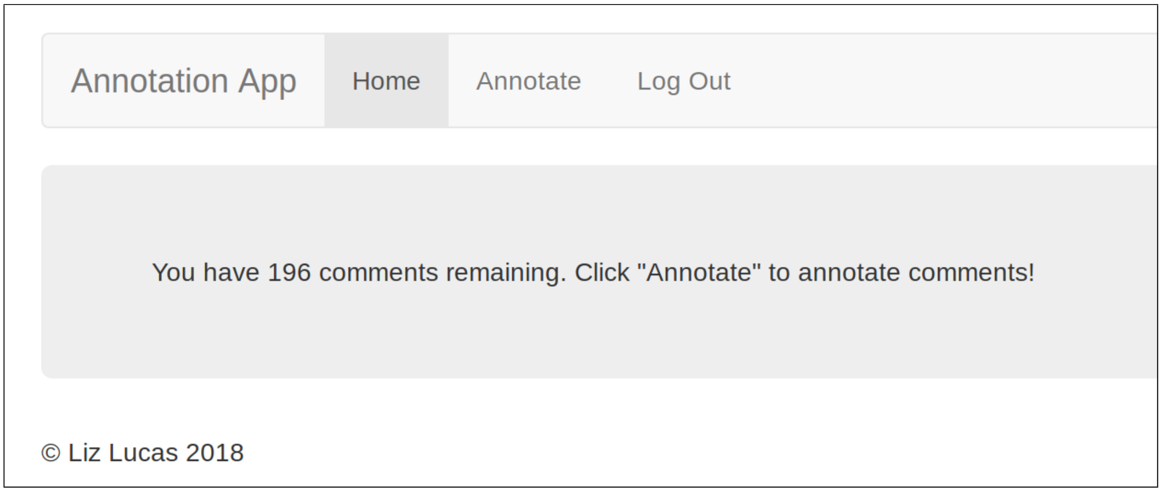
Figure B.3: An additional screen should the user select *not scientific* as a reason.



I'm pretty sure BYU and Science are mutually exclusive

Please explain why this comment is not scientific:

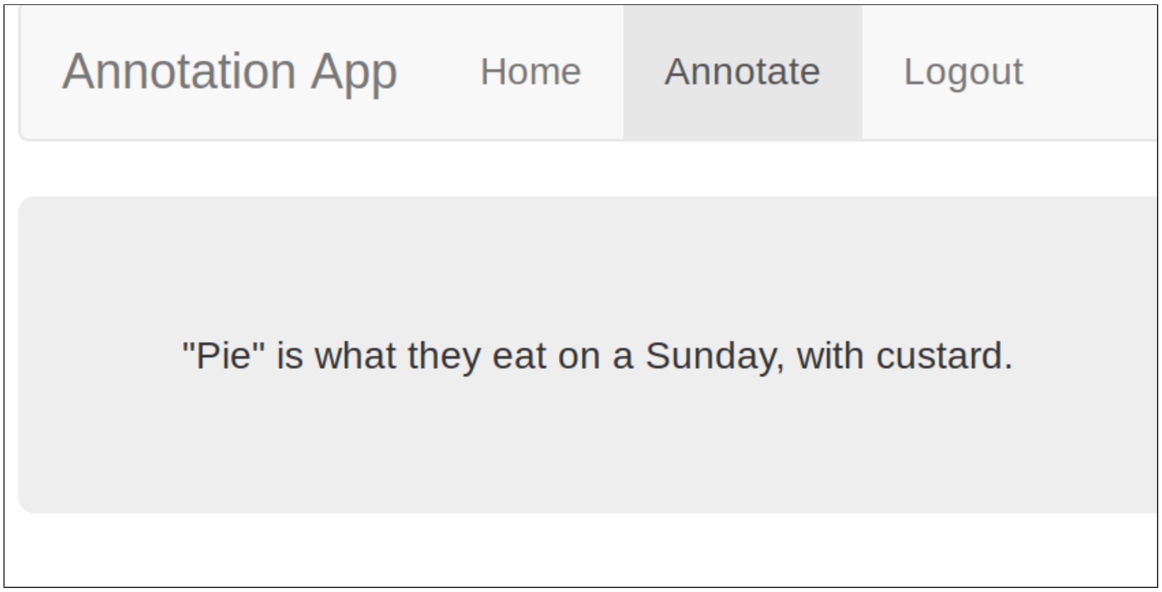
Figure B.4: The home screen for a logged-in user



You have 196 comments remaining. Click "Annotate" to annotate comments!

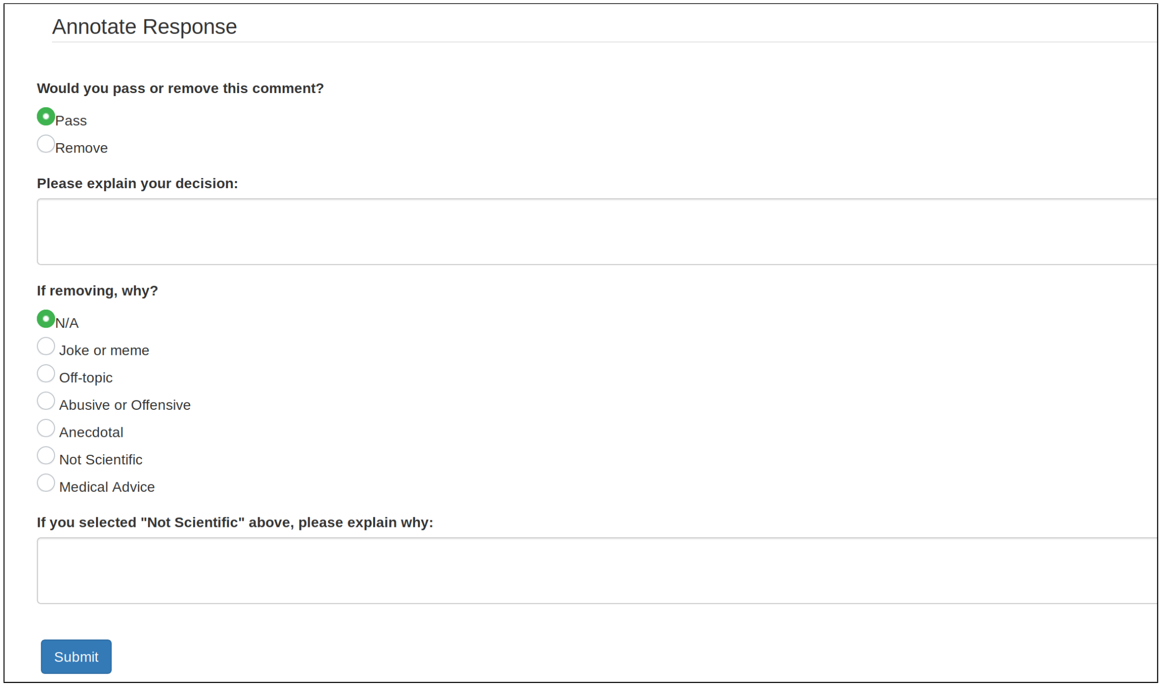
© Liz Lucas 2018

Figure B.5: How the user views comments to be annotated



"Pie" is what they eat on a Sunday, with custard.

Figure B.6: The full form to capture the annotation responses



Annotate Response

Would you pass or remove this comment?

● Pass
○ Remove

Please explain your decision:

If removing, why?

● N/A
○ Joke or meme
○ Off-topic
○ Abusive or Offensive
○ Anecdotal
○ Not Scientific
○ Medical Advice

If you selected "Not Scientific" above, please explain why: