Final Report: A hate speech classifier trained to predict a distribution of ratings

Matthew Edmundson

Abstract— **This project aims to develop and test an alternative methodology for dataset creation informing AI hate speech classifier systems. Information on AI development and training is largely kept private by social media companies that utilise them, including hate speech classifiers that are intended to protect people from being exposed to harmful content. This is problematic as there is little community input nor knowledge on the tools which are control the content they are served online. The methodology proposed by this project attempts to address this by asking people disproportionately targeted by hate speech online to inform the hate speech classifier developed by annotating instances of hate speech to create a dataset according to this project's methodology. Those targeted by hate speech were asked to annotate in subscription to an ethical idea that they have a right to input in this process and they will be more effective at determining what counts as hateful towards members of their own group. As this is a pilot study, practicality meant the scope of this is restricted to people in the Rainbow community classifying Rainbow hate speech comments left online. A substantial process for altering survey design and ethics approval was required on this project, in part due to a more sensitive subject matter and potential to harm for survey participants. The dataset creation methodology developed in this project is intended to improve upon majority rules (gold standard) annotation classification by creating a pilot dataset and methodology which can be used by classifiers for soft label annotation.**

I. INTRODUCTION

Artificial Intelligence (AI) classifiers are used by private technology companies for various purposes. For example, text classifiers are used in conjunction with others such as image classifiers by social media companies like Facebook and Twitter as part of their recommender systems to pair users with content they are likely to want to view. Another use of these classifiers is determining whether content submitted by users follows the companies' community standards and whether it should be removed from the platform. An example of such a classifier is a textual hate speech classifier [1].

Hate speech classifiers work by training a classifier on a corpus of pieces of text, each given a category or 'rating' by human judges. The classifier learns features within the text which are indicative of hate speech or not hate speech. The presence or absence of these features on unseen pieces of text indicate to a classifier whether the text should be classified as hate speech or not.

An issue with how object classifiers typically function is that in creating the training set, the presence of a class is determined through consensus of multiple human judges, i.e., the most popular decision is chosen. This final determination results in the loss of information in the event of conflicting determinations by human judges. Only the label is considered, not the disagreement, which is information that could be used by the classifier to make it more effective. This project proposes a methodology to create a classifier that more precisely measures the hatefulness of comments on a continuous scale of hatefulness rather than categorically hateful and non-hateful. This is done by considering the disagreement between human judges' annotations in the training dataset. Given more precise measurements of comment hatefulness, social media platforms have more granularity in actions they can respond with to new posts and comments appearing on their platform.

II. RELATED WORK

Background research on this project took the form of a literature review. This will be summarised in this section.

A. Existing Dataset Creation Methodologies

'Learning from Disagreement: A Survey' [2] mentions the existing methodologies employed to create datasets used in ML/AI tasks. The classic option is to recruit one individual (typically an expert in the relevant field) to annotate a dataset. One problem with this option in subjective classification tasks is that a dataset outputted by just one individual is subjective. The dataset has a higher potential to be mislabelled. Another problem is that this process is time consuming. Additionally, who defines who is an expert or authority in a field such as hate speech Another option is the gold standard process of annotation [3] in which other expert(s) in the same field offer their input on labels as a separate adjudicating step in the process. This process aids with the problem of subjectivity somewhat and reduces the time required to create datasets but is more expensive. A third option is to crowd-source the annotation process to multiple non-expert individuals through crowd-sourcing platforms and aggregating the answers to output a majority wins label. This process is reportedly more time and financially efficient. Crowd-sourced and aggregated labels are a common methodology but there are issues related to this labelling process. There may not always be 'one correct answer'. For example, in the instance of subjective

This project was supervised by Prof. Ali Knott (primary) and Dr. Andrew Lensen.

classification tasks (such as hate speech classification) classification depends on an individual's subjective outlook and thought process. In the end, the classifier must reduce to a final decision on a comment instance, but preserving information about judge disagreement throughout the training process a more precise final judgement can be made.

B. Alternate Proposed Methodologies

Research into alternate proposed methods focused on crowd annotation types [2]. These can be broken down into four categories:

1. Automatically aggregate crowd-sourced annotations into one label per instance. This method must either assume that a gold standard truth exists and can be found, or, assuming it can be found improves the accuracy of the classifier.
2. Automatically aggregate crowd-sourced annotations into one label per instance but filter out instances with disagreement, i.e., remove instances with conflicting classifications form judges above a threshold. This method assumes a gold standard truth exists but may not always be found.
3. Create a probabilistic distribution from crowd-sourced annotations, producing a soft label. Teach the classifier based on soft labels.
4. A hybrid method in which a classifier is trained using gold labels supplemented by crowd-sourced annotations.

C. Sources of Disagreement Between Human Annotators

Four potential sources for inter-annotator labelling disagreement were presented in the following sources [4, 5]. The first two, errors and interface problems and annotation scheme, are problematic and are required to be minimised in the dataset creation methodology. Errors and interface problems are disagreements resulting from user error i.e., accidentally selecting an unintended label. These errors are reported to be particularly high in NLP tasks with reportedly 15-30% of disagreement resulting from annotator error in certain tasks. Annotation scheme errors are errors resulting from imprecise or overlapping classes in classification schemes. The following two sources of disagreement may not necessarily be negative for classification tasks if the hypothesis of this project is true. Ambiguity is disagreement resulting from genuine ambiguity in a classification task. For example, a sentence can have many meanings depending how it is interpreted. Subjectivity is similar to ambiguity but is distinct in that the meaning of the instance (e.g., sentence) is widely understood by disagreeing annotators but judgements on that meaning may differ. Inter-annotator disagreement because of ambiguity and subjectivity are useful information as they reflect human understanding on the instances. Ambiguity and subjectivity will exist in unseen instances so should be trained into the classifier to make it more generalised, thus more reliable for unseen text.

D. Soft Label Evaluation

The softmax function is reportedly the most effective method for creating soft labels [2]. Uma et al. [2] also found that when training on datasets of sufficient size, with many annotations by high quality annotators per item, probabilistic distributions (soft labels) were most effective in training classifiers. When these prerequisites did not hold, hard label evaluation methods found gold labels achieved the best performance. Soft label evaluation methods found a hybrid of soft labels and gold labels achieved the best performance.

There are different soft label evaluation metrics proposed depending on the measure. For example, to evaluate 'how similar the distribution of labels assigned by the model to an item is to the distribution of judgements produced by an annotator for that item' [2] cross-entropy [6] and Jensen-Shannon divergence [7] were proposed. Cross-entropy was proposed to capture how confident a model is compared to humans and how reasonable its distribution is. Jensen-Shannon divergence was proposed to capture similarity between two probability distributions. An alternate measure to reproducing human judgements is a measure of reproducing inter-annotator agreement given an assumption annotator distribution entropy is a measure of how confusing annotators' find the item. Cosine similarity can be used to produce an entropic similarity measure. Pearson's correlation can be used to produce an entropic correlation metric.

III. DESIGN

This section will cover the design of the survey from a high level, discussing the planning and design of the survey as they relate to the project aims. Justification for more granular survey choices will be discussed in section IV: Implementation.

A. Project Aims

We do not exactly know how social media companies develop and train the AI that perform classifications on their platforms as this information is kept private. This is an issue as we, the public using social media, cannot have a say on the function of a technology that has a large impact on what content we are exposed to online. This is especially problematic when considering AI such as hate speech classifiers that are designed to protect people from being exposed to harmful content. In this project, we propose a methodology that could be employed by social media companies to enhance the precision of classifiers, giving companies a larger range in responses to posts and comments on their platform. This methodology also proposes a way to give the public more input into their function. How the public would have input in their function and is discussed further in section III D.

The overall aim of this project is to develop and evaluate an AI system which classifies hate speech content more precisely than current classifiers used in industry by social media companies. This is proposed to be done by incorporating a measure of hateful content on a continuous scale of 'how hateful is this post/comment?' rather than a discrete classification of hateful and non-hateful. At the project's inception this overall aim was separated into two distinct aims.

The first aim is the creation of a hate speech dataset and dataset creation methodology. This dataset is designed to train classifiers which consider disagreement between human judges. The process of creating the dataset is intended to act as a pilot study. The created alternatively labelled dataset is at much too small a scale to train an effective classifier. Instead, the aim is to develop the methodology of creating an effective dataset incorporating soft labelling (probability distribution labelling). As part of this, a small-scale dataset is created in this project which can undergo analysis. At a larger scale, such as being incorporated in social media classifier systems, the dataset created is intended to be more fit for purpose of training a classifier in which disagreement between multiple human judges is considered.

The second aim is the development of a soft label hate speech classifier which is trained on labels derived from a probability distribution of multiple judges' classifications of seen instances. There is an interesting hypothesis to test here.

We can train two classifiers on the same dataset, using different methods.

- A first classifier can be trained on 'hard labels'. For this classifier, the target output for each content item is given by the winning category identified by annotators. (A one-hot probability distribution can be created, where all the probability mass in the distribution is allocated to this one category.)
- A second classifier can be trained on 'soft labels'. For this classifier, the target output for each content item is the true distribution over annotators' labels, where the probability mass can be distributed over multiple (or all) possible labels.

In both cases, the classifier can be trained using cross-entropy as the loss term, which is the normal way of training classifiers.

We can then evaluate the performance of both classifiers on a test set held out from training. Interestingly, we can evaluate not only for accuracy as to the winning category identified by annotators (which is the standard evaluation metric), but also for disagreement. How well do the classifiers predict the amount of disagreement about a given test item? Both classifiers can make predictions about this. They both produce an output that can be interpreted as a probability distribution: the entropy of this distribution for a given item provides a good measure of the network's uncertainty about this item (and, indirectly, of the disagreement amongst annotators). But the 'soft labels' classifier might be expected to be better at predicting disagreement: because it is explicitly trained to reproduce probability distributions over annotator decisions. This is a hypothesis that could readily be tested once a dataset is created.

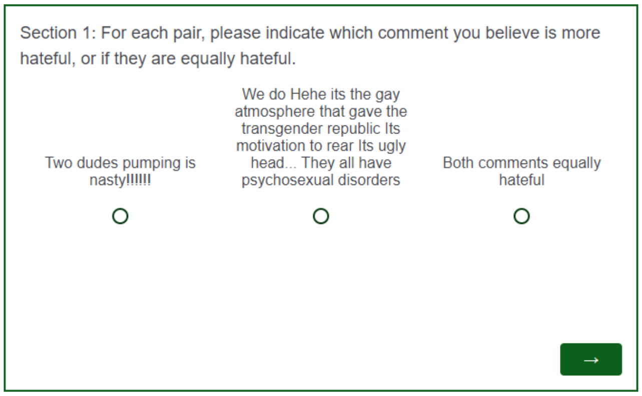
B. Project Aims Change

Over the course of completing this project, it became apparent that that the second aim, development of a soft label classifier, was out of scope. Development, running, and analysis of the dataset and dataset methodology took longer than anticipated in the originally project scope. This was predominately due to the time taken to research and develop multiple iterations of the survey design. Changes to the survey are reflected in the multiple iterations of the ethics application of this project, which is discussed further in section IV B. The care required to ensure the project does its due diligence in caring for participants is warranted. Especially in a project with sensitive subject matter where we are exposing Rainbow individuals to real world hate speech comments about them/a group they identify with. Changes to the survey also occurred to ensure more reliable respondent annotations. This is discussed in section III C.

C. Survey Design

The survey design is dictated by the requirements for the resulting dataset. One requirement is that the dataset captures both discrete and ranking classifications from respondents.

Pairwise comparisons are employed to capture ranking classifications in the first half of the survey. Participants are shown a pair of comments at a time, randomly picked from the full set of 20. During the project the classification options for each comparison changed from a forced choice between two comments to a choice between the two comments and a third 'Both comments equally hateful' option as can be seen in Figure 1.



Section 1: For each pair, please indicate which comment you believe is more hateful, or if they are equally hateful.

Two dudes pumping is nasty!!!!!!

○

We do Hehe its the gay atmosphere that gave the transgender republic Its motivation to rear Its ugly head... They all have psychosexual disorders

○

Both comments equally hateful

○

Figure 1: Survey screenshot - pairwise ranking

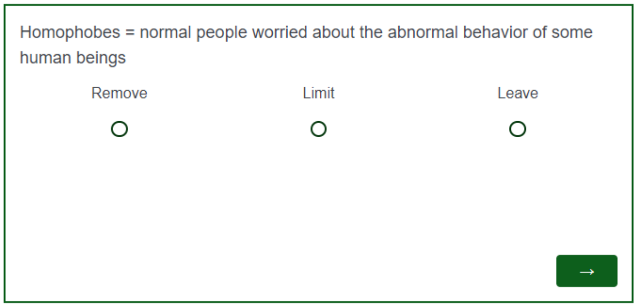
The forced choice was changed to avoid noise in the outputted dataset where respondents picked a comment at random at random if they found them equally hateful. The law of large numbers suggests if the dataset was annotated by enough respondents this noise would average out and be reduced [h]. Besides the fact that any avoidable noise in a dataset should be reduced, if possible, in a pilot project survey with a small scale it is likely the averaging of noise would not occur.

Pairwise comparison was chosen for ranking classification as opposed to a simpler classification such as a

Likert scale for each comment for two reasons. The first is, comparative ratings are reportedly more accurate than absolute ones [8]. It is also suggested in the task of ‘abusive content detection [9]. It is reportedly more difficult to give quantitative assessments to the classification item than simply comparing one item to another. One could also imagine that over the course of completing the survey, participants get more of an idea of the sort of comments being classified and adjust their classification accordingly. This could potentially result in classification variance based on whether comments appeared earlier in the survey or later. By simplifying using pairwise comparison this is avoided.

Simplifying the survey for participants is important as user error or interface problems are reported to result in 15-30% of inter-annotator disagreement in certain NLP annotation tasks [4,5]. These sorts of errors are particularly high in NLP tasks. This is an undesired source of disagreement. We wish sources of disagreement between annotators to be from difference of opinion between annotators. Another desired source of disagreement is from subjectivity in human reading understanding. Different people will have different semantic and sentimental understanding of written text. We wish to capture this to generalise a classifier based on real human understanding.

The discrete categories we ask participants to classify comments on in the second half of the survey are: ‘Leave’, ‘Limit’, and ‘Remove’. An example of a comment classification can be seen in Figure 2. These classifications are based on the respondent’s opinion on what action a social media platform should take in response to the comment. Leave the comment as is, limit its proliferation by down ranking it in the recommender system, or remove it from the platform entirely. By asking respondents to classify based on action we are capturing not just the respondents’ recognition of hate speech, but also the respondents’ attitude towards free speech on social media. By capturing this, participants have more of an impact on the function of the classifier. This is an aim of the project - give the public more input into the function of social media hate speech classifiers.



Homophobes = normal people worried about the abnormal behavior of some human beings

Remove	Limit	Leave
○	○	○

Figure 2: Survey screenshot - discrete categorisation

D. Survey Participants

We endeavoured to have just participants who identify as Rainbow contribute to the project survey. This was done by asking participants to act in good faith on survey advertising material and only participate if they identify as Rainbow. We asked for only Rainbow individuals to participate in this project for two reasons:

1. An ethical subscription to the idea that individuals from groups more commonly targets of hate speech online should have the opportunity to be part of the process of classifier creation.
2. A belief that individuals in a targeted group will be more effective at determining what counts as hateful towards members of their own group.

The targeted group was restricted to just Rainbow as this is an exploratory project. Scope is kept smaller to be more manageable. It is intended if a full-scale study were to be undertaken based on this study that other groups that are commonly recipients of hate speech online would have an additional focus. It is posited in this report that individuals in a targeted group will be more effective at classifying hate speech towards members of their own group. They have a lived experience that can only be felt by part of their community. Further research is required to solidify this hypothesis.

In a full-scale dataset creation process organisers could have individuals who do not identify as part of the targeted group participate. Responses from individuals in the targeted group could then be weighted more heavily to ensure they still impact the classifier process more heavily than others. If this classifier were to be employed on a social media platform, a way to evaluate the weight individuals from targeted groups should have could be done through A/B testing. This idea is discussed more in section VI A. At scale it is intended, given a jurisdiction and harmful content category, each social media company would use the same training set to train classifiers. The training set would be developed outside of the companies semi-publicly. The process of how the training set is created should be shared with the public. The actual contents of the dataset should not be, to avoid adversarial methods to be developed to avoid platform classifiers. The identity of annotators should not be shared to ensure they are not vulnerable to coercion, though these annotators should come from the public. This would function somewhat similarly to the idea of a citizens’ jury where small groups of randomly selected citizens give input on policy decisions by giving recommendations to organisers. This would make the process of defining hate speech on social media more democratic.

IV. IMPLEMENTATION

The design of the survey has been dictated by a combination of requirements of the output dataset and ethical concern for the wellbeing of participants. We have endeavoured to have participants of the survey be people in the Rainbow community. The participants then classify comments chosen from the

English subset of the FRENK open-source social media hate speech dataset available from the Hugging Face repository online [10]. The comments were originally posted on Facebook comments of news article posts about Rainbow topics. These comments were picked to contain varying levels of hate. The level of hate is subjective and is attempted to be more closely determined through running this survey. Though, it was subjectively attempted to find a range of hatefulness in the comments ranging from non-hateful to clearly hateful and/or hateful directed at Rainbow individuals.

A. Survey Platform

The survey was run online on the platform Qualtrics. It was discussed whether it should be run in person, asking participants to complete the survey on the Victoria University campus. This was a consideration was because, during the development of the dataset creation methodology, it was considered whether participants should be able to discuss each comment before classifying. Running the survey in person would also allow for more personalised support for each participant given the sensitive subject matter. In the end running the survey in person was decided against. It is possible that respondents' hate speech opinions would be influenced by the 'loudest voice in the room'. This was not desired as it would be less representative of the multiple participants actual views. It was decided running the survey online would also allow for more participants to respond and from more varied walks of life. Running the survey in person at the university campus would likely result in more homogeneity between respondents. Many would be young adult university students. By running the survey online, ideally, participants demographics would be more varied and more representative of the Rainbow Community. There are limitations to this idea. These are discussed further in section VI: Evaluation. Survey advertisement material was posted online in Rainbow social media groups. It was also posted around the Victoria University Kelburn campus and in S&M's Cocktail Bar – a Rainbow friendly bar in Wellington.

B. Support for participants

Running this survey clearly has ethical implications as we are exposing Rainbow individuals to comments which are hateful to themselves/members of a group they identify as belonging to. Much care has been taken by the researchers to support participants before, during and after they complete the survey. This is reflected in the survey design and the ethics application submitted to Victoria University for this project which is viewable as an appendix to this report – Appendix 1: Ethics Application. It was key participants were able to give informed consent before beginning the survey. To do this, a description of the survey process was displayed to participants before classification began. Participants were also shown two examples of hate speech comments that were attempted to be representative of the level of hate and targets of hate in the survey comments. They are as follows:

- "They don't know there gender!?, That's is mental illness, If you are a girl, then you are If you are a man, then you are How cann't you know your gender. I call it mental illness."
- "Fucking disgusting these bendovers."

Before being shown these examples participants were also provided with descriptions and contact information to general and Rainbow specific mental health support services.

C. Survey comments

The full comments that are being classified by the survey participants are displayed here. The comments are shortened in section V: Analysis to be displayed conveniently. A mistake was made in creating the survey. Comment 5 is dissimilar for the discrete classification portion of the survey to the pairwise ranking section of the survey. Some of the comments were swapped out later on in the development process to include more varied subject Rainbow identities. The survey went live to participants with comment 5 swapped out for the pairwise ranking section and not swapped for the discrete classification section of the survey.

Survey Comments

Comment ID	**Comment**
1	Two dudes pumping is nasty!!!!!!
2	We do Hehe its the gay atmosphere that gave the transgender republic Its motivation to rear Its ugly head... They all have psychosexual disorders
3	DO YOU NOT KNOW THE WICKED WILL NOT INHERIT THE KINGDOM OF GOD? DO NOT BE DECEIVED: NETHER THE SEXUALLY IMMORAL NOR IDOLATER NOR ADULTERERS NOR MALE PROSTITUTES NOR HOMOSEXUAL OFFENDERS NOR THIEVES NOR THE GREEDY NOR DRUNKARDS NOR SLANDERERS NOR SWINDLERS WILL INHERIT THE KINGDOM OF GOD. - 1 Corinthians 6:9-10 NIV } } } *So have fun trying to convince God your dumbass theory :>
4	Who wants their 7 year old daughter to have to take a dump next to a man who 'identifies' as a woman?
5 (Discrete Classification)	you are just so wrong marriage is for procreation, I'll ask you this have you ever seen to 2 bulls make 1 calf no it take 1 cow & and 1 bull you don't comprehend biology.
5 (Pairwise Ranking)	Heidrun that is absolutely a problem. We have designated toilets for males and females. You chose the toilet based on

	your gender not what you want to identify yourself as in your fairytale world.
6	Y isn't there a straight pride march. Homophobia is a pathetic thing but having a parade to celebrate being gay is just as pathetic
7	Are u gay?
8	Being a gay is the biggest disaster in the world
9	Homophobes = normal people worried about the abnormal behavior of some human beings
10	Well if it's not lgbt then it would be a race or someone else not protected by the majority. It will always be one small group or another being attacked. The masses are easier to control when sepearated. Are you on the hate train?
11	He's got AIDS, He's got AIDS, He's got AIDS, He's got AIDS, He's got AIDS,
12	OK, why isn't there a straight pride day march? Probably because non gays have better things to do with their time , like raising their families.
13	I suggest dat d 'HOMOS' should acquire 'common sense', and stop making noise about this absurd idea.. That countries are legalizing 'it' doesn't mean that it is good.
14	IF YOU WERE NOT BORN THROUGH HOMOSEXUALITY...JUST STOP SUPPORTING AND DEFENDING SUCH INSANE PEOPLE
15	Fags
16	Around 3% in the USA but they try to spread their sickness on 97% of the USA population Homo or trans - homo it doesn't matter , that's all the same , these are sick human creatures who lost their own identity and who are agressively pursuing their sickness as some type of "norm " and they should be stopped from doing that
17	It's a disgrace how everyone and everything needs to change according to man's lust, this is a abomination says "GOD THE ALMIGHTY" !
18	There is no excuse for they way they dress. No shame and where is the "Pride" they so proudly tout???
19	Careful or I'll come for you in your dreams with my lesbian powers. You might catch gay!!
20	Stupid People Same Sex Marriage Is just just So Stinky ANd Stupid

V. ANALYSIS

The finished dataset has been downloaded from Qualtrics and loaded into Jupyter notebooks for analysis to be performed. The full raw dataset is viewable as an appendix to this report as Appendix 2: Full Dataset. The analysis notebooks are also viewable as appendices as Appendix 3: Discrete Analysis and Appendix 4: Ranking Analysis.

There are 20 comments in the dataset. At the time of writing this report 35 participants have responded in the pairwise ranking section of the survey. 34 participants have responded in the discrete category judgement section of the survey. The one partial completion of the survey has not been removed as the two sections of the survey are somewhat independent.

A. Discrete Analysis

Winning Judgement Count

Winning Judgement	**Count**
Leave	8
Limit	4
Remove	8

Of the 20 comments in the survey the majority of respondents recommended the social media platform response of leave on the platform as is for 8 comments, limit the proliferation of 4 comments, and remove 8 comments completely off the platform.

Winning Judgements

Comment ID	**Comment**	**Winning Judgement**
1	Two dud...	Leave
2	We do H...	Remove
3	DO YOU ...	Remove
4	Who wan...	Remove
5	you are...	Leave
6	Y isn't...	Leave
7	Are u gay?	Leave
8	Being a...	Remove
9	Homopho...	Leave
10	Well if...	Leave
11	He's go...	Remove
12	OK, why...	Limit
13	I sugge...	Limit
14	IF YOU ...	Remove
15	Fags	Remove
16	Around ...	Remove
17	It's a ...	Limit
18	There i...	Leave
19	Careful...	Leave
20	Stupid ...	Limit

Discrete Probability Distribution

A discrete probability distribution over the normalised frequencies can be calculated on the three classifications for each comment. We can see an example of this plotted in Figure 3. All plotted distributions are viewable in Appendix 5: Plotted Probability Distributions.

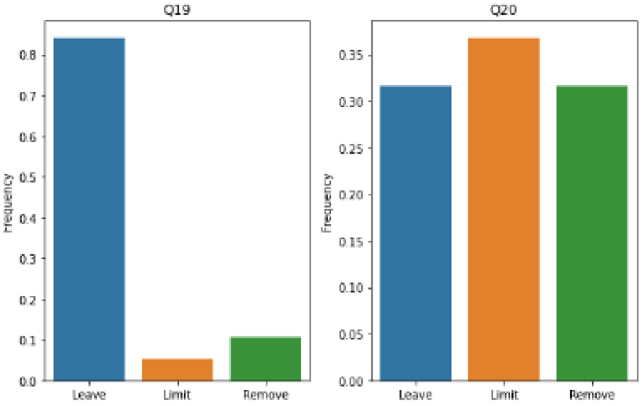


Figure 3: Normalised frequency over categories

- 19: Careful or I'll come for you in your dreams with my lesbian powers. You might catch gay!!
- 20: Stupid People Same Sex Marriage Is just just So Stinky ANd Stupid

Figure 3 illustrates that there are differences in the inter-annotator disagreement between comments. For example, the vast majority of respondents classified comment 19 as 'Leave' – that it should be left as is on the platform. For comment 19 classification there is low inter-annotator disagreement. In contrast, for comment 20 the winning judgement is 'Limit' but there is a close margin to the other categories. For comment 20 there is high inter-annotator disagreement.

Comment Classification Entropy

From these distributions we can calculate a measure of entropy for each comment. This entropy measure can act as a measure of confidence for our classifications. High entropy means high inter-annotator disagreement and less certainty on classification. Low entropy means low disagreement and more classification certainty.

The formula used to calculate entropy is as follows:

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

Figure 4: Cross-entropy formula [11].

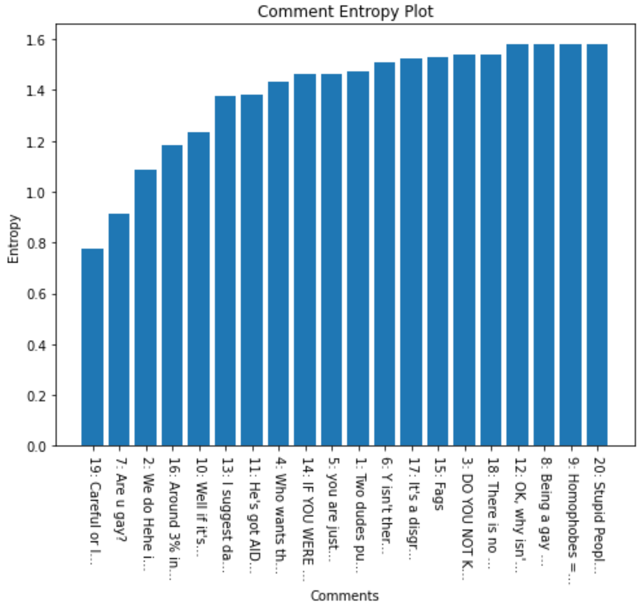


Figure 5: Comments by entropy

By viewing Figure 5, graphing of comments by entropy, we can see that comment 19 has the lowest disagreement whereas comments 12, 8, 9, and 20 have a similarly have a similarly high level of disagreement.

We can average entropy by classification to see which winning judgements have most agreement – Figure 6.

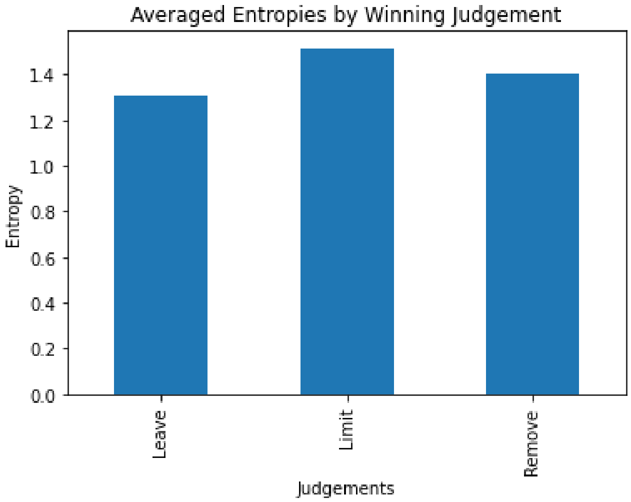


Figure 6: Averaged entropies by Winning Judgement

Average Winning Judgement Entropy			
Winning Judgement	**Averaged Entropy**	**Standard Deviation**	**Variance**
Leave	1.310804	0.308149	0.094956
Limit	1.516085	0.095884	0.009193
Remove	1.399607	0.177266	0.031423

We find, on average, there is lowest disagreement on comments where most respondents chose to leave the comment. There is slightly more disagreement when most respondents chose to remove the comment completely. There is most disagreement when most respondents thought a comment should be limited on the platform.

Order of Confidence

We can order comments on a scale of increasing annotator confidence by ordering by entropy. The more certain the winning judgement, the higher the comment is in the ranking.

Ordering of Confidence			
Comment ID	**Comment**	**Winning Judgement**	**Entropy**
19	Careful or I...	Leave	0.774243
7	Are u gay?	Leave	0.913283
2	We do Hehe i...	Remove	1.086988
16	Around 3% in...	Remove	1.181291
10	Well if it's...	Leave	1.236160
13	I suggest da...	Limit	1.377963
11	He's got AID...	Remove	1.383808
4	Who wants th...	Remove	1.432983
14	IF YOU WERE ...	Remove	1.461838
5	you are just...	Leave	1.461838
1	Two dudes pu...	Leave	1.471354
6	Y isn't ther...	Leave	1.509028
17	It's a disgr...	Limit	1.524317
15	Fags	Remove	1.529428
3	DO YOU NOT K...	Remove	1.539491
18	There is no ...	Leave	1.539491
12	OK, why isn'...	Limit	1.581031
8	Being a gay ...	Remove	1.581031
9	Homophobes =...	Leave	1.581031
20	Stupid Peopl...	Limit	1.581031

Ordering of Hatefulness

We can order all the comments in the dataset by winning judgement and entropy. If we say that 'Leave' classifications are less hateful than 'Limit' classifications' which are in turn less hateful than 'Remove' classifications, this ordering can act as a scale of hatefulness. Comments classified as 'Leave' are ranked in order of lower entropy as the lower the entropy the more certain we can be of the positive classification. For the same reason, 'Remove' comments are ranked in order of higher entropy. The higher the entropy the less certain we are of a negative classification. For comments with a majority 'Limit' classification we can rank by the normalised 'Leave' judgement value minus the normalised 'Remove' judgement value descending. If we were to order by entropy alone on these 'Limit' classifications, we are not capturing the values of 'Leave' and 'Remove', only disagreement between all classifications.

Ordering of Hatefulness			
Comment ID	**Comment**	**Winning Judgement**	**Entropy**
19	Careful or I...	Leave	0.774243
7	Are u gay?	Leave	0.913283
10	Well if it's...	Leave	1.086988
5	you are just...	Leave	1.181291
1	Two dudes pu...	Leave	1.236160
6	Y isn't ther...	Leave	1.377963
18	There is no ...	Leave	1.383808
9	Homophobes =...	Leave	1.432983
12	OK, why isn'...	Limit	1.461838
17	It's a disgr...	Limit	1.461838
20	Stupid Peopl...	Limit	1.471354
13	I suggest da...	Limit	1.509028
8	Being a gay ...	Remove	1.524317
3	DO YOU NOT K...	Remove	1.529428
15	Fags	Remove	1.539491
14	IF YOU WERE ...	Remove	1.539491
4	Who wants th...	Remove	1.581031
11	He's got AID...	Remove	1.581031
16	Around 3% in...	Remove	1.581031
2	We do Hehe i...	Remove	1.581031

B. Pairwise Ranking Analysis

In the pairwise ranking section of the survey participants were presented with two random comments from the dataset and asked to choose whether one comment was more hateful than the either or if they thought both comments were equally hateful. The dataset outputted shows respondents' opinions on the relative hatefulness of a series of two comments.

We can create a pairwise ranking leader board of these responses, incrementing by one for a win for one comment over another and adding 0.5 to both comments if they were determined to be equally hateful [12]. We can then run a Bradley-Terry model on the dataset to simulate the outcomes of pair comparisons that were not explicitly ranked on one another [13, 14]. If we wished to compare every item against every other item, we would require 190 pair comparisons to get just one unique pairwise comparison each: N(N-1)/2, N=20. In this exploratory project respondents gave a total of 332 pairwise comparisons. One to two individual pairwise comparisons are an insufficient number to rank comments accurately. The Bradley-Terry model allows to get a prediction of ranked scores when not all comments are compared given calculated probabilities from the pairwise comparisons we do have. The model setup can also be viewed in Appendix 4: Ranking Analysis. The model scores each individual comment after simulating 'matches' between pairs based on the inputted probabilities. The model converged after 19 iterations and calculated the following scores for the comments:

Ordering by Hatefulness Score		
Comment ID	**Comment**	**Hatefulness Score**
7	Are u gay?	0.31
19	Careful or I...	0.54
10	Well if it's...	0.99
20	Stupid Peopl...	2.12
15	Fags	2.60
18	There is no ...	2.67
1	Two dudes pu...	2.85
9	Homophobes =...	3.60
12	OK, why isn'...	3.63
6	Y isn't ther...	3.85
5	Heidrun that...	4.85
11	He's got AID...	5.04
14	IF YOU WERE ...	5.21
17	It's a disgr...	6.46
2	We do Hehe i...	6.52
8	Being a gay ...	7.24
13	I suggest da...	7.47
4	Who wants th...	7.69
3	DO YOU NOT K...	8.69
16	Around 3% in...	17.69

C. Joint Set Analysis

In the section 'Ordering of Hatefulness' we order comments first by winning judgement then entropy measure (with the entropy order function depending on winning judgement). We can also combine discrete judgement and pairwise ranking datasets to order comments by winning judgement and the Bradley-Terry hatefulness score.

Unfortunately, as mentioned in section IV: Implementation, the discrete dataset and ranking dataset mistakenly contain one comment each that does not appear in the other dataset (Comment 5). For joint classification, we will remove these comments from the dataset.

Ordering by Winning Judgement and Hatefulness Score			
Comment ID	**Comment**	**Winning Judgement**	**Hatefulness Score**
7	Are you gay?	Leave	0.31
19	Careful or I…	Leave	0.54
10	Well if it's…	Leave	0.99
18	There is no …	Leave	2.67
1	Two dudes pu…	Leave	2.85
9	Homophobes =…	Leave	3.60
6	Y isn't ther…	Leave	3.85
20	Stupid Peopl…	Limit	2.12
12	OK, why isn'…	Limit	3.63
17	It's a disgr…	Limit	6.46
13	I suggest da…	Limit	7.47
15	Fags	Remove	2.60
11	He's got AID…	Remove	5.04
14	IF YOU WERE …	Remove	5.21
2	We do Hehe i…	Remove	6.52
8	Being a gay …	Remove	7.24
4	Who wants th…	Remove	7.69
3	DO YOU NOT K…	Remove	8.69
16	Around 3% in…	Remove	17.69

In the ranking above we can see that the hatefulness scores of comments follow the general trend of judgement categories, i.e., 'Remove' comments have a higher hatefulness score than 'Limit' which in turn are higher than 'Leave'. There is cross over though, where the highest hatefulness scores in each category are lower than many of the scores in the other categories. This is especially visible where comment 13 in the 'Limit' category has a hatefulness score of 7.47 and comment 15 in the 'Remove' category has a hatefulness score of 2.60, significantly lower. Potential sources for this noise will be discussed in section VI: Evaluation.

We could create a hate speech classifier to be employed on a social media platform trained on a full-scale dataset created in the fashion of this survey. We have several options on how this could be done given the dataset. We train on a dataset of the discrete winning judgement and associated entropy, or by pure hatefulness score as a regression task, or by a combination of discrete winning judgement and hatefulness score. Another alternative is to, first, create a ranking of comments by winning judgement and hatefulness score. We then employ both a classification model and regression model. The classifier using this dataset could leave new comments alone on the platform that are classified in the 'Leave' classification. Using this we remove new comments completely from the platform that are classified as 'Remove'. A regression model could calculate entropy on new comments. This could be an NLP transformer with a regression head, given we want to include text embeddings, using entropy as a loss training function. We can use a combination of entropy and hatefulness score for new comments to down rank on the platform more or less that fall in that are classified as 'Limit'. Comments such as these are typically referred to as 'borderline content' This is a hybrid methodology, incorporating gold standard classification and soft labelling. Comments are down ranked less if there is high entropy, these are comments where there is more disagreement. This is to preserve some conflicting opinions the platform to keep discussions alive as a matter of free speech.

It may be that a social media platform wants to create cut-off points for leaving, limiting, and removing comments at different points depending on internal policy. The cut off points for removing comments could be at a lower or higher level based on hatefulness score and entropy. For example, as a basic measure, the ranking score below has been calculated from the normalised hatefulness score plus entropy. A platform could use this measure for more granularity.

Ordering by Judgement and Calculated Ranking Score			
Comment ID	**Comment**	**Winning Judgement**	**Ranking Score**
19	Careful or I…	Leave	0.000000
7	Are you gay?	Leave	0.106688
10	Well if it's…	Leave	0.401282
1	Two dudes pu…	Leave	0.668526
18	There is no …	Leave	0.718212
6	Y isn't ther…	Leave	0.738420
9	Homophobes =…	Leave	0.788619
20	Stupid Peopl…	Limit	0.731518
13	I suggest da…	Limit	0.769153
12	OK, why isn'…	Limit	0.789777
17	It's a disgr…	Limit	0.851827
2	We do Hehe i…	Remove	0.490658
11	He's got AID…	Remove	0.680257
15	Fags	Remove	0.707148
14	IF YOU WERE …	Remove	0.751670
4	Who wants th…	Remove	0.823371
8	Being a gay …	Remove	0.929059
3	DO YOU NOT K…	Remove	0.950477
16	Around 3% in…	Remove	1.000000

In due course one may be able to analyse disagreement specifically for the pairwise ranking task. This is an avenue for further research as it is out of scope for this project.

VI. EVALUATION

In this section we will evaluate the created dataset and the dataset creation methodology in relation to the aims of the project.

A. Classifier Evaluation

As mentioned in section III B. the design, implementation, and analysis of the dataset took longer than was initially anticipated at the project's inception. Evaluation of the classifier is therefore more hypothetical currently. Instead, we will discuss what future evaluation would look like. It would be possible to evaluate with a soft label classifier by sourcing a 'majority wins' labelled hate speech dataset from online that has preserved the frequency of the other judgements. Majority wins labels determine classification through the highest frequency of multiple judges' annotations. A probability distribution could be calculated over the frequency of the winning and non-winning comment classifications. The product of this would be a soft-label dataset. A classifier could be trained on this and evaluated to evaluate the efficacy of the sort of dataset developed in this project. This evaluation is not completely corollary to the dataset methodology proposed in this project as the sourced dataset would be co-opted to create soft labels not developed from scratch. The discrete categorisation would be similar, but the ranking regression would not be present.

It is not possible for us to evaluate a classifier with a full-scale dataset as proposed in this project. The scale required for such a dataset is far out of the scope of this project. It would be possible, however, for a social media company to test such a dataset. A social media company could evaluate this dataset and classifier methodology through A/B testing on their platform. They would deploy incorporate the hybrid classifier in the recommender systems of certain subsets of their user base. They could then evaluate its effectiveness through explicitly asking users for feedback on their perceptions of the hatefulness of posts they see. The platform could also measure user complaints on the platform to hate speech between testing subsets. Similarly, the platform could measure feedback in complaints of free speech suppression in the two subsets. The balance between these two could further inform how much hatefulness score and entropy are weighted in calculating comment ranking scores in the dataset which informs the classifier, and subsequently, how much borderline content is downranked. The subset of users the social media platform test this moderation schemes in could come from varied communities on the platform and differences between them could be similarly evaluated. If this is implemented it would be an evaluation based upon the health of a social media platform as a whole when it comes to balancing promoting productive discourse and protecting individuals from hate speech.

B. Inter-Annotator Disagreement Measures

In the proposed methodology inter-annotator disagreement is calculated through cross-entropy. Another way at getting at inter-annotator is through Fleiss' kappa statistic, or Krippendorff's alpha [15, 16]. These methods are not suggested for this task as they do not provide a cross-entropy measure that a model can train on.

C. Dataset Values Evaluation

In section V: Analysis one can see in the table Ordering of Confidence that there are multiple entropy values over winning judgements that are the same. For example, comment 14 and 15 both have the entropy value of 1.539491. This is a problem because we wish to have a dataset with more granularity in the measures between comments. The more granularity one has, the more precisely one can down rank items within a recommender system. The origin of the similar values is the relatively small number of annotators on the dataset. The more annotators you have, the less likely it is the frequency of classifications will be the same between comments, meaning it will be less likely to have the same entropy values.

We can also see in the table Ordering by Hatefulness Score that there is cross over between categories where the highest hatefulness scores in each category are lower than many of the scores in the other categories. Would we not expect that the hatefulness score of a comment discretely classified in the 'Limit' category would have a lower hatefulness score than a comment classified in the 'Remove' category? A potential source of this error is again, the relatively small number of annotations on the dataset. The variance on hatefulness scores estimated by the Bradley-Terry model will likely be reduced as the frequency of pair comparisons increase.

D. Comment Context

An aspect of that is missing from the design of the survey is that comments are presented alone, without the context of the post they were originally replying to on Facebook. It's possible that knowing the context of the comment changes whether it is considered as hate speech. For example, consider a comment in response to a post: "welcome to the world of the gays". Is this a welcoming comment to a community or a degrading, sarcastic comment on a post which contains negative sentiment? This nuance is currently missed by the design of the survey. Subjectivity is a factor in the classification of the comments but commenters intent through implicit writing tone can be made clearer to annotators with further context of the original post. If research in this area were to be continued, this is an avenue that can be further explored.

E. Online Survey Limitations

The survey was ran online utilizing the platform Qualtrics. The main reason this was done was to allow the highest number of participants from more varied demographics within the Rainbow community to participate. An obvious problem with this is people who do not feel comfortable or do not have ready access to devices to go online are less likely to participate. Future work could explore avenues to solve this problem.

There also exists a self-selection bias in the survey. This is reduced as the survey offers supermarket vouchers to attract volunteers who otherwise might not have but it is still present. If the survey is run at a larger scale self-selection bias would also be reduced by increasing diversity and reducing the impact of outliers.

VII. CONCLUSIONS AND FUTURE WORK

This project proposes a method to improve upon social media hate speech classifiers in which information from disagreements between annotators is preserved. This is a hybrid methodology that incorporates discrete gold standard labelling, continuous scores for regression synthesized from pairwise ranking, and soft labels created from probability distributions from crowd annotations over discrete categories. A methodology to create a dataset more fit for purpose for training a classifier is also proposed. This is done through gathering participant annotations in both discrete and pairwise ranking format. It is also proposed how social media companies might incorporate a classification system such as this into their platforms and evaluate its effectiveness – by A/B testing in different communities on their platform. The true test of the efficacy of the hate speech classification system is the health of the platform in promoting free speech in constructive discussions and protecting users from hate speech.

As mentioned in section III: Design, it is possible to evaluate a hard label classifier against a soft label classifier trained on a majority wins co-opted dataset. This was not completed in this project as it was determined it was out of scope as the project progressed. If research on this topic were to be continued, this would be an avenue for further discovery. If research is continued, the impact of including comment context, such as the original social media post, can also be explored.

REFERENCES

[1] M Hampson, "Combating Hate Speech Online With AI: It analyzes the context of social media posts to accurately detect hate speech", https://spectrum.ieee.org/ai-versus-online-hate-speech, (Accessed: 1/6/2023).

[2] A. Uma et al., "Learning from Disagreement: A Survey," Journal of Artificial Intelligence Research, vol. 72, pp. 1385-1470, Dec. 2021, DOI: https://doi.org/10.1613/jair.1.12752.

[3] C. Cheng. H. Asi. J. Duchi. "How many labelers do you have? A closer look at gold-standard labels," DOI: https://doi.org/10.48550/arXiv.2206.12041.

[4] A. Nedoluzhko. et al., 2016, "Coreference in Prague Czech-English Dependency Treebank," In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 169– 176, Portorož, Slovenia. European Language Resources Association (ELRA).

[5] S. Pradhan et al., 2012, "CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes," In Proceedings of Joint Conference on EMNLP and CoNLL - Shared Task, pp. 1–40, Jeju Island, Korea, Association for Computational Linguistics.

[6] J. C. Peterson, et al., 2019. "Human uncertainty makes classification more robust," In Proceedings of International Conference on Computer Vision (ICCV), pp. 9616–9625, DOI: https://dl.acm.org/doi/10.5555/2391181.2391183.

[7] J. Lin, "Divergence Measures Based on the Shannon Entropy," IEEE Transactions on Information Theory, vol. 37, no. 1, pp. 145-151, 1991.

[8] R. D. Goffin. J. M. Olson, 2011, "Is It All Relative? Comparative Judgements and the Possible Improvement of Self-Ratings and Ratings of Others," Perspect Psychol Sci., vol. 6, 1, pp. 48-60, Available: https://pubmed.ncbi.nlm.nih.gov/26162115/, DOI: 10.1177/1745691610393521

[9] S. Kiritchenko. I Nejadgholi, 2020, "Towards Ethics by Design in Online Abusive Content Detection," arXiv preprint arXiv:2010.14952, Available: https://arxiv.org/pdf/2010.14952.pdf

[10] N. Ljubešić. D. Fišer. T. Erjavec., 2019, "The FRENK Datasets of Socially Unacceptable Discourse in Slovene and English," https://huggingface.co/datasets/classla/FRENK-hate-en, Available: https://arxiv.org/abs/1906.02045

[11] S.Kwiatkowshi, 2018, “Entropy is a measure of uncertainty,” Towards Data Science, Available: https://towardsdatascience.com/entropy-is-a-measure-of-uncertainty-e2c000301c2c

[12] “Pairwise comparison method,” 1000 minds, https://www.1000minds.com/decision-making/pairwise-comparison

[13] D. R. Hunter, 2004, “MM algorithms for generalised Bradley-Terry models,” Ann. Statist., vol, 32, 1, pp. 384-406, Available: https://projecteuclid.org/journals/annals-of-statistics/volume-32/issue-1/MM-algorithms-for-generalized-Bradley-Terry-models/10.1214/aos/1079120141.full, DOI: https://doi.org/10.1214/aos/1079120141

[14] S. Mohammad, 2021, “Code: How Bradley Terry Model Works,” Kaggle, Available: https://www.kaggle.com/code/shaz13/code-how-bradley-terry-model-works

[15] J. L. Fleiss, 1971, "Measuring nominal scale agreement among many raters.," Psychological Bulletin, vol. 76, no. 5 pp. 378–382, DOI: https://doi.org/10.1037/h0031619

[16] A. Zapf. et al., 2016, “Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate?” BMC Med Res Methodol, vp;. 16, no. 93, Avaiable: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-016-0200-9 , DOI: https://doi.org/10.1186/s12874-016-0200-9