PteroDACTyl: Predicting Disagreement About Comment Toxicity

Junsol Kim
Department of Sociology
University of Chicago
junsol@uchicago.edu

Irene Tang
Department of Linguistics
University of Chicago
itang1@uchicago.edu

Fiona Lee
Computational Social Science
University of Chicago
mlee26@uchicago.edu

Abstract

Flagging toxic content in online media is not a one-size-fits-all task. While a growing line of research seeks to develop models to detect, classify, and address toxic content, internet users with different backgrounds, experiences, and perspectives often perceive the same piece of content differently—one population viewing a certain comment as toxic while another deeming it acceptable. To address this limitation, we built two models (one using DistilBERT as its base, and one using RoBERTa) that predict not only which social media comments appear toxic to different demographics, but also how much disagreement will arise about whether a comment is toxic. We found that the three models performed comparably in terms of accuracy on the task of predicting whether a comment is not toxic, subjectively toxic, or blatantly toxic. The model with RoBERTa base performed slightly better than the other two in terms of F-1 scores.

1 Introduction

Toxic content and offensive language are prevalent in social media and online platforms, appearing in the form of threats, hate and harassment comments, identity attacks, slurs, and other provocative language. A growing area of research seeks to develop models to detect, classify, and address such online content. However, this is not a one-size-fits-all task. Internet users with different backgrounds, experiences, and perspectives often perceive the same piece of content differently—one population viewing a certain comment as toxic while another deeming it acceptable (Kumar et al., 2021; Aroyo et al., 2019).

Considering this divergence in perspective, Kumar et al. (2021) conducted a large-scale survey (N=17,280) on how people from different backgrounds perceive social media comments sampled from Twitter, Reddit, and 4chan. They found that individuals and demographic groups who have personally or historically been a target of harassment (such as people identifying as LGBTQ+, young adults, etc.) are more likely to flag comments as toxic, compared to those who have not experienced harassment. This subjective nature of content toxicity poses a challenge to content moderation in online platforms: it underscores the need for diverse raters in ground-truth data labeling, and it allows for gray areas where an online content can harm a fraction of users without explicitly violating any moderation policies. These survey results call for moderation that goes beyond one-size-fits-all classifications.

After unveiling this subjectivity in perception, Kumar et al. (2021) proposed the idea of tuning classifiers to personalize the threshold of what is considered harassment. Their idea was to use a moving threshold that can be set higher or lower to match an individual or demographic group's perspective on toxicity. Compared to Jigsaw (2021)'s Perspective API (a free API that uses machine learning to identify toxic comments) and Instagram (2019)'s comment nudge (a feature that nudges a user if they are about to post a comment similar to those that have been flagged in the past) baselines, Kumar et al.'s personalized tuning model increased accuracy by 86% per individual and 22% per demographic group with respect to flagging whether or not a comment appears toxic to the rater. This customized model addresses the fact that personal experiences impact one's perception of toxicity.

Our project addresses a limitation to Kumar et al. (2021)'s study: while Kumar et al. revealed that perceptual differences exist amongst individuals and demographic groups, they did not discern which linguistic characteristics make certain comments seem toxic to some but not all people. This limits our understanding about the nature of the discordance regarding toxicity. To address this limitation, we built models that predict not only which social media comments appear toxic to different demographics, but also how much disagreement will arise about whether a comment is toxic.

2 Our Contributions

This project builds upon Kumar et al. (2021)'s work, predicting not only which social media comments appear toxic to different populations, but also how much disagreement there will be across populations about whether a comment is toxic. Our contributions are threefold.

First, we present SPinOPs (Subjective Perception in Online Places), a new dataset of comments and their subjectivity classification based on the human judgements survey data collected by Kumar et al. (2021). In this new dataset, each surveyed comment is matched with one of three labels: 1) not toxic, 2) subjectively toxic, and 3) blatantly toxic. Criteria for how labels were assigned are described in 3.2.

Second, we present PteroDACTyl, a pair of pretrained models (DistilBERT and RoBERTA) that we fine-tuned on our new dataset to predict judgement subjectivity. DistilBERT was chosen for its compact, yet high-performing, capabilities (Sanh et al., 2019).[1] RoBERTa was chosen for its improved training procedure that improved BERT's performance on several downstream tasks (Liu et al., 2019). BERT-based models have demonstrated state-of-the-art performance in this realm of identifying toxic content—such as in the Kaggle multilingual toxicity classification challenge, where Lee (2020)'s first-place submission obtained AUC scores up to 0.95, and in the NLP industry, where Jigsaw (2021)'s Perspective API obtained AUC scores up to 0.97. As such, we expected that our two BERT-based models will also effectively capture linguistic patterns underlying disagreements on toxicity.

Third, we examined what makes a comment get flagged as toxic in the eyes of both humans and computational models. For human judgements, we manually inspected the data for linguistic patterns that differentiate subjectively toxic comments from blatantly toxic comments, as well as for patterns of speech that are particularly inflammatory to different demographic populations (i.e. patterns that are predictive of human disagreement). And for our PteroDACTyl models, we examined whether it picks up on any of these same linguistic patterns to use as heuristics, or whether it conducts its classification based on other features that are non-intuitive to humans.

These contributions should be useful to content moderators in online communities where people with diverse backgrounds interact and often conflict, and will provide insight about linguistic patterns that are integral to disagreement.

3 Dataset

3.1 Original Dataset

The data that we used in this project is based on survey data provided by Kumar et al. (2021). Kumar et al.'s dataset consists of 107,620 social media comments amassed from Twitter, Reddit, and 4chan. Each comment was judged by five different human annotators[2] on many factors, including whether it includes toxic language, profanity, threats, identity attacks, insults, and sexual harassment. Judgements from all five annotators are included in the dataset, along with the respective annotator's demographics (e.g. gender, age, race/ethnicity, LGBTQ+ status, religion, political attitude, family structure, etc). A sample datapoint is provided in Appendix A.1.

A total of 17,280 annotators participated. 52% were female, 47% male, and 2% unknown/nonbinary. 71% were White, 12% Black or African American, 6% Asian, 3% Hispanic, and 8% multiracial/other/unknown. 40% are between ages 25-34, 25% between 35-44, 13% between 45-54, 12% between 18-24, and 7% between 55-64. 41% have a bachelor's degree, 20% some college but no degree, 15% a Master's degree or higher, 11% an Associate's degree, and 9% a high school diploma. 83% are heterosexual, 11% bisexual, 4% homosexual, 3% undisclosed/other. 77% have personally seen toxic content and 29% have personally been targeted. 40% are politically liberal, 27% conservative, 27% independent, and 6% undisclosed/other. 90% use social media, 55% follow news media, 86% use video media platforms, and 54% use fo-

[1] DistilBERT retains 97% of BERT's language understanding capabilities but runs 60% faster and uses only 40% of its size (Sanh et al., 2019).

[2] Kumar et al. (2021) made the decision to solicit five ratings per comment based on a pilot survey in which 10 unique participants rated each pilot stimulus for toxicity on a five-point Likert scale ranging from "Not at all toxic" to "Extremely toxic." In measuring how many judgements it takes for each comment's ratings to converge to its average toxicity score, they found that five judgements was optimal for balancing this convergence with the monetary cost of soliciting a judgement.

rum discussion platforms. 52% were parents, 47% non-parents, and 1% undisclosed/other.

In aggregate, 52% of the individual judgement instances (i.e. one annotator's judgement on one comment) deemed that the corresponding comment was *0-not toxic*, 19% as *1-slightly toxic*, 13% as *2-moderately toxic*, 9% as *3-very toxic*, and 6% as *4-extremely toxic*. 15% of judgements claimed that the comment included profanity, 7% a threat, 14% an identity attack, 20% an insult, and 4% a sexual harassment (with some comments possessing multiple of these qualities).

From these distribution statistics, it appears that although an effort was made to solicit judgements from a diverse population, the proportion of respondents unfortunately skewed toward the majority demographic. Most of the respondents were White, millennial-aged, college-educated, and heterosexual. A more-even distribution appeared among political views, but still there were significantly more liberal respondents than respondents from other parties. Gender and parenthood status were rather balanced. While this distribution of survey respondents could plausibly be representative of the adult media-consuming population as a whole, a balanced distribution of demographics and perspectives would have been more expedient for training models on the task of predicting how diverse populations would perceive potentially-toxic comments.

The saving grace among the homogeneity is the dataset's profound sample size. Despite majority demographics being over-represented in proportion, the sheer quantity of surveyed individuals and comments provides a decent quantity of example judgements to work with, even for under-represented demographics.

3.2 Our Dataset: SPinOPs

Using Kumar et al. (2021)'s human judgements survey data, we created SPinOPs (Subjective Perception in Online Places), a new dataset that matches each comment to one of three labels: subjectively toxic, blatantly toxic, not toxic. The labeling rationale is as follows. If the set of human judgements on a comment's toxicity level had high between-individual variance (defined below), then this means that there was a lot of disagreement about whether that comment was toxic, and so that comment received the label of *subjectively toxic*. On the other hand, if a comment solicited low between-individual variance in how humans judged it, then that means that there was much agreement about how toxic that comment was. For these comments, if it had a high mean toxicity score, we labeled it as *blatantly toxic*; and if it had a low mean toxicity score, we labeled it as *not toxic*.

Between-individual variance for any given comment ($V_{individual}$) was calculated according to Equation 1. Given a comment x, each annotator's toxicity rating of that comment is denoted as x_i, the mean toxicity rating across all annotators who rated that comment is denoted as $\overline{x}$, and the number of annotators who rated that comment is denoted as n.

$$V_{individual} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} \qquad (1)$$

Between-group variance for any given comment (V_{group}) was calculated according to Equation 2.[3] Given a comment x, the number of different groups being compared is denoted as g, each group's mean toxicity rating of that comment is denoted as $\overline{x}_i$, the global mean toxicity rating across all groups is denoted as $\overline{x}$, and the number of annotators in each group who rated that comment is denoted as n_i.

$$V_{group} = \sqrt{\frac{\sum_{i=1}^{g}(\overline{x}_i - \overline{x})^2 n_i}{g-1}} \qquad (2)$$

A value of $V = 0$ indicates perfect agreement among the individuals or groups being compared, and higher values of V indicate higher levels of disagreement.

We set the threshold for what is considered "high variance" to be the third quartile of between-individual variances. (In this dataset, this value was 0.89.) Thus, if a comment's between-individual variance in toxicity ratings is larger than this threshold, then we label it as *subjectively toxic*. Else, since its between-individual variance is lower than this threshold, we look to its mean toxicity ratings to assign a label. If its mean toxicity rating is 2 or greater,[4] then we label it as *blatantly toxic*; and if its average toxicity rating is below 2, then we label it as *not toxic*.

[3] This measure of between-group disagreement is known to be better in estimating between-group variance in sparse and continuous data compared to other measures (such as Fleiss Kappa), and it has been extensively used in previous NLP studies, especially studies based on ANOVA models (Kumar et al., 2021; Cohen, 1968; Wooldridge, 2015).

[4] 2 = moderately toxic. This is the center value of the Likert scale, which runs from 0-4.

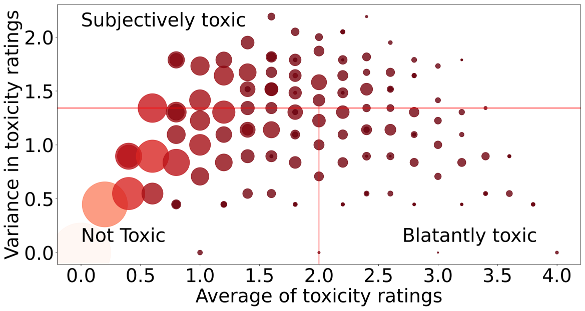

Figure 1: Distribution of mean and variance in the ratings. Lighter, larger circles correspond indicate greater quantities of comments receiving that mean/variance combination.

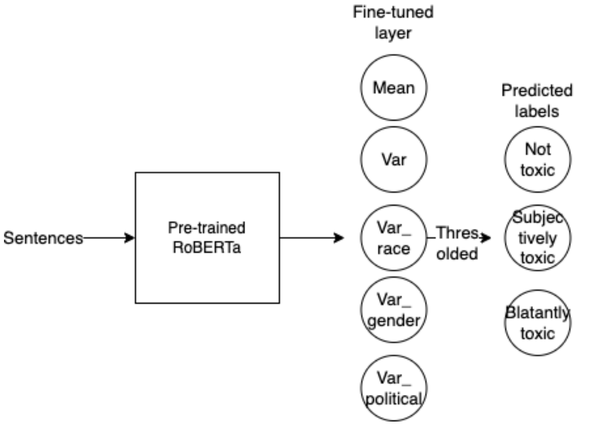

Figure 2: Model architecture.

Under this methodology, 68.89% of the comments in SPinOPs were labeled as not toxic, 22.40% as subjectively toxic, and 8.71% as blatantly toxic. The distribution is visualized as in Figure 1.

4 Models

4.1 Model specifications

We trained two BERT-based models, collectively called PteroDACTyl (Predicting Disagreement About Comment Toxicity), to predict five toxicity ratings associated with a given comment. Specifically, we use RoBERTa-base and DistilBERT-base model implemented in Python transformers package.

Figure 3 describes the architecture of our models. The input comments are first subjected to preprocessing and to the addition of special instance markers ([CLS], [SEP], etc.). The pre-processed input is then tokenized using the DistilBERT or RoBERTa tokenizers. The sequence of vectors in contextual representations from the these BERT models is then fed to our fine-tuning layer (linear layer) which predicts five toxicity scores: 1) average toxicity rating, 2) between-individual variance in toxicity ratings, 3) between-race variance in toxicity ratings, 4) between-gender variance in toxicity ratings, and 5) between-political variance in toxicity ratings. After generating predictions for the five qualities stated above, a simple arithmetic calculation classifies the comment as either not toxic, subjectively toxic, or blatantly toxic as described in Section 3.2.

The BERT-based models' performance was compared against a third baseline bag-of-words (BOW) model. In the baseline BOW model, unigram binary features (whether words are included in a sentence (=1) or not (=0)) are used as features to support vector regression models that predict five toxicity scores.

Based on the grid-search technique (LaValle et al., 2004), we used the following hyper-parameter values which maximize model performance in the validation dataset: number of epochs $= 2$ (best out of 2, 4, 6), learning rate $= 10^{-6}$ (best out of 10^{-6} and 10^{-5}), dropout rate $= 0.2$ (best out of 0.1, 0.2, 0.3). We used mean squared error (MSE) as the loss function to optimize the fine-tuned layer (training loss $= 0.4383$, validation loss $= 0.4257$).

4.2 Model evaluation

In this section, we evaluate the PteroDACTyl models' performance against that of a simple bag-of-words model.

To evaluate our model, we split our dataset into training (80%), validation (10%), and test (10%) datasets. The validation dataset is used to optimize hyper-parameters of the model, and the test dataset is used to evaluate the models. The following results are based on the test dataset.

Figure 3 presents the prediction performance of the fine-tuned layer that predicts five toxicity scores. We found that 1) average toxicity rating is strongly correlated to the predicted scores in RoBERTa model (Pearson's r= .672; Mean squared error (MSE)= .374). Similarly, 2) Between-individual variance (r= .399, MSE= .270), 3) between-race variance (r= .232, MSE= .547), 4) between-gender variance (r= .221, MSE= .547), and 5) between-political variance (r= .302, MSE= .434) in toxicity ratings are significantly correlated with the predicted scores. The predictive power of RoBERTa-based model was higher than BOW and

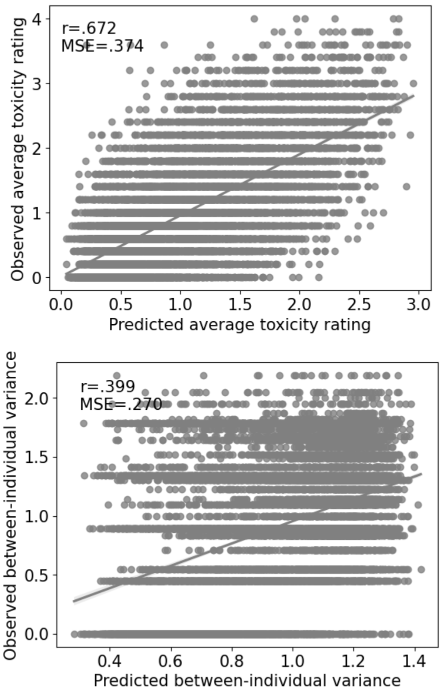

Figure 3: Predictive performance of fine-tuned RoBERTa model.

DistilBERT model.

After that, we categorized comments into three bins based on predicted toxicity scores: 1) not toxic, 2) subjectively toxic, and 3) blatantly toxic as shown in the Section 3.2. Since it is a multi-class prediction task, we utilized macro- and weighted-averaged F1-score, Precision, Recall, and Accuracy [5]. The results are shown in Table 1.

The RoBERTa model showed the highest performance in predicting labels (weighted F-1 score=.621). The RoBERTa and DistilBERT models were shown to effectively capture the differences between not toxic comments, subjectively toxic, and blatantly toxic comments However, the BoW regression model could not discern not ambiguous differences among these comments.

4.3 Error Analysis

BOW (Bag-of-Words). BOW had a hard time recognizing toxicity. It classified 95% of comments in the test set as not toxic, compared with the 68.89% ground-truth human judgements percentage–suggesting that BOW missed most of the patterns in comments that human annotators rated as toxic, such as race-related subjectively toxic comments [6], and blatantly toxic comments containing unspeakable racial slurs. For subjectively toxic predictions on not toxic comments, BOW captured the ambiguity aspect of utterances with missing context [7]. The only mis-classified instance that BOW predicted as blatantly toxic could not provide any insights or patterns for such classification.

DistilBERT. Although with a higher accuracy in predicting toxicity and variances, DistilBERT also missed a number of instances that human annotators rated as subjectively toxic, which include race and religion-related comments as well as short phrases with missing contexts. Blatantly toxic comments that DistilBERT missed (classified as not toxic or subjectively toxic) tend to be sentences that contain implicit and mild derogation that may require interpretation with semantics and contexts[8]. On the other hand, DistilBERT was sensitive to unspeakable racial slurs and cursing, while these mis-classified instances are perceived subjectively toxic to human annotators. Although these comments may be acceptable to the human annotators randomly assigned, they may be intolerable to others. The mismatch between DistilBERT's predictions and human annotations may suggest what human annotators have overlooked due to different perspectives and tolerance, while DistilBERT was able to capture and maintain the bottom line standards of toxicity.

RoBERTa. RoBERTa generated predictions closer to the human annotated toxicity of comments. Although it similarly missed (classified as not toxic) a number of comments that human annotators perceive blatantly or subjectively toxic, most of them are short statements without contexts. Looking at the mis-classifications that RoBERTa classified as blatantly toxic, unspeakable racial slurs and explicit cursing are not the only elements that trigger blatant toxicity prediction, while it also considered insulting phrases and words that could be source of toxicity Example: "the stormfag conspiracy is why the Holocaust is being ridiculed!". In addition, comments that RoBERTa mis-classified as subjectively toxic have similar characteristics: race or sexual-related, and short statements without meaningful contexts. Some of

[5] The macro-averaged metrics are the arithmetic mean of per-class metrics. The weighted-averaged metrics are weighted mean of per-class F1 scores based on per-class proportion.

[6] Example: "Why is blacks also in the Jew parentheses?"

[7] Example: "That sucks dude, best of luck."

[8] Example: "you're too stupid to see that though"

Evaluation Metrics	BoW Regression	DistilBERT-base	RoBERTa-base
F-1 score (weighted)	0.2937	0.6017	**0.6213**
Precision (weighted)	0.3034	0.6243	**0.6303**
Recall (weighted)	0.3542	0.6905	**0.7108**
F-1 score (macro)	0.3056	0.3853	**0.4249**
Precision (macro)	0.4135	0.4753	**0.5233**
Recall (macro)	0.3547	0.4125	**0.4539**
Accuracy	0.7194	0.7122	0.7108

Table 1: Model evaluation

these comments that human annotators perceived as not toxic seem to be "over-killed" by RoBERTa Example: "You obviously don't know much about Muslims.", while this could reveal the dilemma of placing a lower threshold of toxicity for better protection, versus lifting the threshold and let people decide and filter toxic contents.

5 Linguistic Patterns Analysis

In this section, we perform descriptive analysis in identifying linguistic patterns that humans and models might be using to flag toxicity, especially patterns that appear particularly inflammatory to different demographic populations. For the SPinOPs dataset and each of the three models (DistilBERT, RoBERTa, BOW), we selected the top 50 not toxic, subjectively toxic, and blatantly toxic comments (as described in Section 3.2). We investigated these three categories for linguistic patterns (including words and phrases, use of non-alphanumeric characters, grammar, spellings) that might be indicative of toxicity scores.

5.1 Human Judgements

Looking at the comments in the SPinOPs dataset, a few patterns stood out.

5.1.1 Between-individual variance.

Among the not toxic comments, there were a lot of instances of phrases and short statements describing the writer's emotions and feelings about non-sensitive topics, and sometimes ending with a sequence of emojis. No common theme emerged regarding the nature of the topics.

Blatantly toxic comments overwhelmingly tended to contain unspeakable racial slurs or race-based derogation, often in conjunction with cursing, and especially targeting Black and Jewish populations. To a lesser extent, they contained gender-based derogation, overt sexual solicitations, or suggestions for the reader to kill themselves.

By contrast, the subjectively toxic comments have less obvious infractions. Many include references to race or nationality, but these appear in the form of assertions of factual statements rather than as expressions of personal opinions about a particular demographic.[9] The missing surrounding context is likely the most important deciding factor in these cases. A number of mildly sexually suggestive innuendos appear too, which might be considered acceptable to some but not annotators, or may appear innocuous to the naive—both types of annotators would assign it a low toxicity score, but for different reasons. The ratings here do not distinguish between raters believing that there is no offense, and raters not noticing the offense.

5.1.2 Between-demographics variance

We also looked at the top 50 comments that were subjectively toxic across three demographic groups: race, gender, and political affiliation.

Race. Across racial groups, a number of comments referenced a particular race, ethnicity, or culture, usually as an unheated expression of opinion that is not necessarily derogatory.[10] There were also several instances of sexual content, mostly (but not always) factual assertions about events rather than opinions about a particular individual or group.[11]

Gender. Across genders, more instances of sexual content appeared, especially following a pornography theme. This wasn't the only (or even majority) content, but there weren't other reliable patterns to distinguish.

[9] Example: "The suspects are described as black males"

[10] Example: Because Communism is a failed political ideology

[11] Example: Did you know that 12 year olds masturbate? What's the harm in it?

Political affiliation. In a similar vein, across political affiliations, the top comments did not demonstrate a clear pattern in common phrases or topics.

We refrain from making generalizations about subjectivity across demographics because distinguishing patterns did not always clearly stand out, and we were even unable to locate the locus of offense in roughly half of the supposedly most-subjective comments.

5.2 Model Heuristics

In this section, we examine whether the two PteroDACTyl models (DistilBERT-based, RoBERTa-based) and the bag-of-words baseline model picks up on any of the same linguistic patterns as humans to use as heuristics, or whether they conduct their classification based on other features that are non-intuitive to humans.

5.2.1 Between-individual variance

BOW (Bag-of-Words). Not toxic comments, surprisingly, contain several racial discriminating instances that include unspeakable racial slurs and cursing, and references to politics. The 50 most blatantly comments predicted by BOW show a more consistent pattern with our finding in SPinOPs. A great number of comments include cursing and derogation regarding multiple races and religions, but many instances are short phrases or incomplete sentences that have no contexts. Unlike in the original SPinOPs dataset, top 50 subjectively toxic comments are more similar to class 2, containing even more occurrences of unspeakable racial slurs, cursing, and short statements about negative personal feelings and complaints. With poor ability in predicting subjectively and blatantly toxic comments, BOW failed to detect words and phrases that are perceived highly toxic by human annotators, and it did not capture the patterns that trigger disagreement on individual level.

DistilBERT. DistilBERT was able to classify similar linguistic styles to not toxic and blatantly toxic classes as human annotators did. The top 50 not toxic comments predicted by DistilBERT include multiple instances that describe positive personal feelings and end with emoticons. The top 50 blatantly toxic comments contain explicit racial insults targeting Black, Jewish, and other racial ethnic populations, associated with derogatory terms and cursing. By contrast, DistilBERT demonstrated a different pattern in classifying subjectively toxic comments. Most of the top 50 subjectively toxic comments are sexual suggestive contents, discussions and opinions on sexual harassment, sexual assaults, and sexual orientation. A small number of political references and opinions regarding political figures also appear. Moreover, these subjectively toxic comments are relatively long and complete than in the other two classes. Compared with human annotations on SPinOPs data set, DistilBERT focused on comments around sexual-related topics, while human annotators would have diverse perspectives over several topics, such as race, politics, and sexual suggestive contents.

RoBERTa. RoBERTa made similar classification decisions as DistilBERT. The top 50 not toxic comments are mostly instances describing positive personal feelings politely. Almost no cursing or derogation exist in these comments. The top 50 blatantly toxic comments contain hate comments and derogatory phrases targeting particular race, ethnicity, and religions as well as homosexuals, and multiple instances also appeared in DistilBERT's predictions, indicating similar prediction ability in classifying blatantly toxic comments. For the top 50 subjectively toxic comments, RoBERTa also focused on sexual-related comments, specifically sexual assaults of children and teenagers. Overall, RoBERTa performed similar classification as DistilBERT, and they both successfully captured the linguistic patterns of not toxic and blatantly toxic comments, while only focused on sexual-related topics for subjectively toxic comments.

5.2.2 Between-demographics variance

All three models performed comparatively with respect to between-demographics variance.

Race. These comments contain a large number of derogatory phrases regarding several racial, ethnic, gender, and sexual orientation groups. Compared with comments selected by individual-level disagreement, these instances selected using variance between races are more closely related to racial topics. However, multiple instances are short phrases or incomplete sentences that do not target a certain person; instead, they are descriptive statements[12] that contain names of racial groups.

Gender. These comments are mostly about sexual topics: sexual suggestive contents, statements regarding sexuality and sexual orientations, and a small number of references to race. Despite high density of sexual terms in these comments, many of them are not used for insulting or derogatory purposes, which could be one potential reason of their highly disagreed toxicity from different genders.

Political affiliation. These comments include discussions and opinions on racial inequality, rights and privileges. A number of instances contain sexual suggestive contents and derogatory phrases associated with race and gender. The high disagreement across political affiliations seem to root from different perspectives of regulations and policies regarding racial groups, along with cursing and disrespectful phrases in the utterances.

Overall, variances in demographics appear in relevant comments. Variances between race tend to root from the ambiguity or non-targeting nature of the comments, while variances in gender and political affiliations are associated with discussion and personal opinions. Although these comments are relevant with each demographic category, we observed that gender and race are often used in derogatory forms in insulting and hate comments, and utterances attacking particular political figures and parties would frequently associate with sexual assaults.

6 Conclusion

In this paper, we presented three contributions to online content moderation. First, we created SPinOPS, a dataset describing how much disagreement occurs across individuals (in general) about how toxic an online comment is. Second, we trained PteroDACTyl, a DistilBERT-based computational model that predicts toxicity ratings, as well how much disagreement about a given model's toxicity rating would occur across race, gender, and political demographics. Finally, we discussed linguistic patterns that could plausibly be predictive of disagreement.

These contributions should be useful to online content moderators in understanding the subjective nature of human perception on toxic content. There is rarely a consensus about what should be considered toxic content—people can have higher or lower tolerances for being offended, and different topics of discussion strike people differently. There are also diverging opinions about what sorts of content should be allowed to appear in media—some would say that any content that appears offensive to even one target should be prohibited, whereas others would be okay viewing offensive language that does not personally target them, and yet others stand for unfiltered liberty of expression at the cost of being offended. PteroDACTyl provides a solution for this sort of demand for customization that media platforms could take into consideration when implementing their content moderation.

7 Future Work

Subjectivity labels (i.e. {not, subjectively, blatantly} toxic) in the SPinOPs dataset were assigned based on the respective comment's between-individual variance in toxicity ratings, which were taken from Kumar et al. (2021)'s survey. An issue arises in that this survey did not control for what sorts of individuals got assigned to any given comment. There was no balance in whether a comment received judgements from a homogeneous or a highly-diverse set of raters; assignments were made by random chance. And with each comment only receiving five annotators' judgements, between-individual variance is highly sensitive to who those five annotators are. Future versions of SPinOPs should address the source data's randomness in annotator assignment.

That being said, we see enormous potential to scale up the size of SPinOPs using semi-supervised learning algorithms. Given the glossy amount of labeled subjectivity data that we already have, and the ease of obtaining more unlabeled instances of online media in today's ripe digital age, there should be a low barrier of entry to correctly infer subjectivity labels for new media data, and then use these newly-labeled data as additional training examples to inductively feed back in to further improve said classification model.

As for PteroDACTyl, one area of further investigation is to regulate the demographic distribution of the survey data that we use to train the models. As discussed in Section 3.2, the dataset that PteroDACTyl trained on was more demographically homogeneous than ideal. Yet given the magnificent

[12] Example: The suspects are described as black males

size of this dataset, there is potential to extract subset of the data such that an even distribution of demographics is represented, while still maintaining a hefty number of example judgements to work with. How would a model that is trained on this smaller, truly-diverse subset perform against PteroDACTyl?

Acknowledgements

We thank Professor Chenhao Tan for his feedback. We also thank the authors behind Kumar et al. (2021) for providing us with their large-scale survey data upon which we built this project.

References

Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. 2019. Crowdsourcing subjective tasks: the case study of understanding toxicity in online discussions. In *Companion proceedings of the 2019 world wide web conference*, pages 1100–1105.

Jacob Cohen. 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Instagram. 2019. Our progress on leading the fight against online bullying.

Google Jigsaw. 2021. Perspective api.

Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing toxic content classification for a diversity of perspectives. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 299–318.

Steven M LaValle, Michael S Branicky, and Stephen R Lindemann. 2004. On the relationship between classical grid search and probabilistic roadmaps. *The International Journal of Robotics Research*, 23(7-8):673–692.

Chun Ming Lee. 2020. Jigsaw multilingual toxic comment classification: 1st place solution overview.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Jeffrey M Wooldridge. 2015. *Introductory econometrics: A modern approach*. Cengage learning.

A Appendix

A.1 Sample datapoint from the original dataset

Table 2: One sample datapoint from Kumar et al. (2021)'s large-scale survey. The table is truncated after reporting the first (of five) annotator's judgements in order to reduce redundancy.

Feature	Value
comment	this is QUINN you DUMBASS
source	twitter
perspective_score	0.90911716
comment_id	1135_5
toxic_score_0	0
is_profane_0	false
is_threat_0	false
is_identity_attack_0	false
is_insult_0	false
is_sexual_harassment_0	false
gender_0	Male
race_0	White
technology_impact_0	Very positive
uses_media_social_0	true
uses_media_news_0	true
uses_media_video_0	true
uses_media_forums_0	false
personally_seen_toxic_content_0	true
personally_been_target_0	false
identify_as_transgender_0	No
toxic_comments_problem_0	Rarely a problem
education_0	Bachelor's degree in college (4-year)
age_range_0	45 - 54
lgbtq_status_0	Heterosexual
political_affilation_0	Conservative
is_parent_0	Yes
religion_important_0	Very important
fine_to_see_online_0	This is fine for me to see
remove_from_online_0	This comment should be allowed
toxic_score_1	2
is_profane_1	false
is_threat_1	false
is_identity$_{a}ttack$_1	false
is_insult_1	true
...	...