PteroDACTyl: Predicting Disagreement About Comment Toxicity

Junsol Kim [1] Irene Tang [2] Fiona Lee [3]

[1]Department of Sociology [2]Department of Linguistics [3]Computational Social Science

Introduction

Addressing toxic and offensive online content is not a one-size-fits-all task!

- People with different backgrounds perceive the same piece of content differently—one population viewing a certain comment as toxic while another deeming it acceptable [4, 1].
- Has been a target of harassment in the past ⇒more likely to flag comments as toxic.

Kumar et al. (2021): Tuned classifiers to personalize the threshold of what is considered harassment.

- Moving threshold that can be set higher or lower to match an individual or demographic group's perspective on toxicity.
- 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, compared to previous baselines [3, 2]

One limitation: Kumar et al. did not discern which linguistic characteristics make certain comments seem toxic to some but not all people.

To address this limitation, we built a model that predicts not only which social media comments appear toxic to different demographics, but also how much disagreement will arise about whether a comment is toxic.

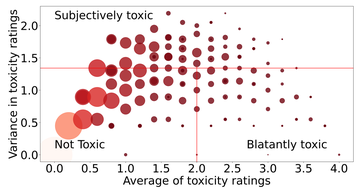
Our Contributions

- **SPinOPs:** A new dataset of comments and their subjectivity classification based on the human judgements survey data collected by Kumar et al. (2021).
- **PteroDACTyl:** A pre-trained RoBERTa-based model that we fine-tuned on our new dataset to predict judgement subjectivity
- **Analysis of linguistic patterns:** An examination of what makes a comment get flagged as toxic in the eyes of both humans and computational models.

Contribution 1: SPinOPs Dataset

Created a new dataset on disagreement about online comment toxicity based on Kumar and colleagues (2021)'s human judgements survey data. Our dataset consists of 107,620 social media comments from Reddit, Twitter, and 4chan rated by multiple annotators. Each comment was matched with one of three labels: 1) not toxic, 2) subjectively toxic, and 3) blatantly toxic.

- **Subjectively toxic (68.89%) =** High between-annotator variance
- **Blatantly toxic (22.40%) =** Low between-annotator variance, high mean toxicity score
- **Not toxic (8.71%) =** Low between-annotator variance, low mean toxicity score



Contribution 2: PteroDACTyl Model

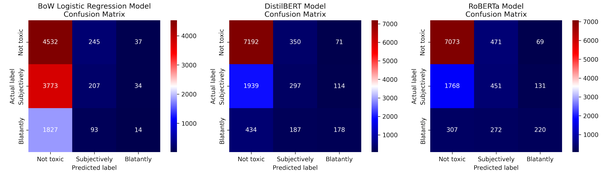
Model design

- RoBERTa-based text classification model to predict inter-annotator disagreement about online comment toxicity. The model discerns (1) not toxic, (2) subjectively toxic, and (3) blatantly toxic comments with AUROC of 75.65% (macro-averaged AUROC across different labels).

Model evaluation

- Training (80%), validation (10%), and test (10%) sample. To contextualize the model performance, we train the baseline bag-of-words logistic regression model and BERT-based models.
- **Bag-of-words logistic regression model:** AUROC of 49.55%. Using unigram binary features, the model could not discern not toxic and subjectively toxic comments well. Out of 10,132 comments predicted not to be toxic, only 44.73% are not toxic.
- **DistilBERT-base & RoBERTa-base model:** AUROC of 70.70% up to 75.65%. These models effectively capture the differences between not toxic comments, subjectively toxic, and blatantly toxic comments compared to the baseline bag-of-word model.

Evaluation Metrics	BoW Logistic Reg	DistilBERT-base	RoBERTa-base
AUROC (macro)	0.4955	0.7124	**0.7565**
F-1 score (macro)	0.3083	0.4434	**0.4865**
Accuracy	0.7295	0.7124	0.7196



Predicting disagreement across demographics

- Considering that the disagreement frequently occurs across demographics, we train another RoBERTa-based model that predicts disagreement between race, gender, and political group.
- Between-group disagreement about comment toxicity is measured as follows. Each toxicity rating is denoted as x, the global mean is denoted as $\bar{x}$, the group mean is denoted as $\bar{x_c}$ bar, and the number of groups is denoted as g:

$$V_{group} = \sqrt{\frac{\sum_{c=1}^{g}(\bar{x}_c - \bar{x})n_c}{g-1}}$$

- The next section presents the analyses on the model outcomes: top-50 comments that are likely to trigger disagreement across demographics.

Future Work

1. Address the source data's randomness in annotator assignment—affects SPinOPs.
2. Scale up the size of SPinOPs using semi-supervised learning algorithms.
3. Regulate the demographic distribution of the survey data that we use to train PteroDACTyl.

Contribution 3: Linguistic Patterns Analysis

Top 50 most-blatantly toxic comments (largest mean toxicity scores) and most-subjectively toxic comments (largest standard deviation in inter-annotator judgements and across demographics)

Blatantly toxic comments

- **Manual:** Contained unspeakable racial slurs or race-based derogation in conjunction with cursing, especially targeting Black populations, gender-based derogation, overt sexual solicitations, or suggestions for the reader to kill themselves.
- **Models:** Mainly contained racial slurs and cursing targeting Black populations.

Subjectively toxic comments

- **Manual:** Less obvious infractions; references to race or nationality in the form of assertions of factual statements rather than as expressions of personal opinions about a particular demographic; mildly sexually suggestive innuendos.
- **Models:** Topics regarding sexuality, sexism, races, criminality, and sexual suggestive contents.

Across race

- **Manual:** Comments referencing a particular race or culture, as an unheated expression of opinion that is not necessarily derogatory. Several instances of sexual content, mostly factual assertions about events rather than opinions about a particular individual or group.
- **Models:** A number of references to races, sexuality, and politics; meaningless phrases and short sentences without contexts.

Across gender

- **Manual:** More instances of sexual content appeared, following a pornography theme. Other contents show no reliable patterns to distinguish.
- **Models:** Topics regarding sexuality and sexual suggestive content; references to races limited to white and black populations. Many instances contained emoticons and hashtags.

Across political affiliation

- **Manual:** No discernible or consistent pattern across comments.
- **Models:** Mainly contained discussions on sexuality and sexual suggestive content; a small number of references to races and politics.

References

[1] Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. 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, 2019.

[2] Instagram. Our progress on leading the fight against online bullying, 2019.

[3] Google Jigsaw. Perspective api, 2021.

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

Acknowledgements

We thank Professor Chenhao Tan for his feedback on this project.