Detecting Harmful Online Conversational Content towards LGBTQIA+ Individuals

Jamell Dacon, Harry Shomer, Shaylynn Crum-Dacon and Jiliang Tang
Michigan State University
daconjam, shomerha, crumshay, tangjili}@msu.edu

Motivation

Queer people often rely on the sanctity of online spaces to escape offline abuse; however, social media remains a hostile, exclusive, restrictive, and controlling environment for gender and sexual orientation, race-related topics, and LGBTQIA+ individuals, activists and allies. and may be distressing for some readers.

To address the above problem, we attempt to detect several forms of toxicity in comments geared towards the LGBTQIA+ community.

Dataset & Annotation

We adapt the queerness (gender/sexual orientation) dimension from REDDITBIAS[1] (see Table 1).

We use Detoxify[2], a model capable of detecting different types of toxicities in text (see Table 2).

We employ Amazon Mechanical Turk (AMT) annotators to manually rate 1000 randomly sampled comments to measure the effectiveness of the toxicity classifier.

We first investigate the overall model performance for the binary classification task in Table 5 using both the macro and weighted F1-score.

We later evaluate the model performance for the multi-label classification task. This task aims to predict 6 distinct harmful comment labels in Table 6, respectively.

We investigate the overall model performance for the multi-label classification task. This task aims to predict 6 distinct harmful comment labels in Table 6, respectively.

Experimental Results

In Figure 1, we illustrate a correlation matrix between labels to determine heavily correlated labels. For example, toxicity vs. identity attack, and insult vs. obscene.

Table 1: Queerness examples comments from REDDITBIAS.

Table 2: Automated labeled queerness (shortened) example comments from Table 1 using Detoxify.

Table 3: Harmful and non-harmful comment counts w.r.t each label for a total of 9930 comments.

Table 4: Macro and weighted F1-score results on all labels for the classes Harmful (1) and Non-harmful (0) into their respective labels (i.e., toxicity (T), severe toxicity (S), obscene (O), threat (Th), insult (In), and identity attack (Id)), and their statistics.

Table 5: Macro and weighted F1-score only on Toxicity label.

In Figure 1, we illustrate a correlation matrix between labels to determine heavily correlated labels. For example, toxicity vs. identity attack, and insult vs. obscene.

Table 6: F1-score results on all labels for the classes Harmful (1) and Non-harmful (0) into their respective labels (i.e., toxicity (T), severe toxicity (S), obscene (O), threat (Th), insult (In), and identity attack (Id)), and their statistics.

Discussion

Now, the question arises, “Why does BERT outperform HateBERT, a model designed specifically for detecting hate speech?”

- Difference in tasks.
- Representation of harmful LGBTQIA+ language in both REDDITBIAS[1] vs. RAL-E[3] datasets.
- BERT allows for a more flexible set of initial parameters.

Future work:

Addressing imbalance issue – Oversampling, sample weighting, resampling, threshold moving, etc.

Limits & Ethical Considerations

- Our work highlights the need for impactful natural language technologies to readily identify harmful online content to minimize further.
- We encourage NLP practitioners to prioritize collecting a set of training data with several examples to detect harmful online content.
- Words associated with swearing or profanity in a text such as “lol bro you gay as fuck lmao” will be classified as toxic, regardless of the tone or the intent of the author.

References

[1] Soumya Barik, Aamer Larueh, Pranali Vaila, and Goran Glavaš. 2021. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics.

[2] Laura Hans and Unityn team. 2020. Detoxify. Github: https://github.com/unityn/detoxify.

[3] Tommaso Caselli, Valerio Basile, Jelena Mitrovic, and Michael Grefenstette. 2020. Marbert: Retraining BERT for abusive language detection in English. CoRR, abs/2010.12472.