Hate Speech Classification

Arnav Kumar Behera, Vedanta Mohapatra

October 2022

1 Introduction

In this assignment, we classified the scraped tweets from Twitter as toxic or non-toxic. The code for the classification can be found at Google Colab Code. We have used the following toxicity classification models and our datasets:

1. toxic-bert
2. Gooogle Perspective API

The assignment aimed to check where these two models diverged and obtain interesting insights from them. All the required .csv files are present in the CSV_files directory.

2 Agreement of the Models

Both the models were used to get the toxicity scores of the tweets from two of the topics - **Asian Hate Crime**, and **Feminism**, and then to make the classification binary, we used a threshold of 0.5 and classified them as non-toxic(0) or toxic(1).

Sometimes, the API calls failed for the Google Perspective API, so some tweets (less than 10) were not classified by it, so these were ignored in the agreement calculations and further analysis.

The following results were found:

1. Asian Hate Crime
   - Krippendorff's Alpha: 0.619425
   - Raw Agreement: 0.881407
2. Feminism
   - Krippendorff's Alpha: 0.600783
   - Raw Agreement: 0.946573

These results though not excellent, were satisfactory, to say the least. Tweets where these models disagreed and tabulated. Now, random samples were extracted from these tables and observed.

Now, a good manual annotation would have required some framing of guidelines and following them diligently. But, we thought of experimenting a little and created some Google Forms (Asian Hate Crime Response Form, Feminism Response Form) and asked few people to fill these to classify tweets as toxic, or non-toxic and also mention if they required context, including us.

Now, most of the tweets that the models disagreed upon were also not agreed upon by the human annotators. Now, both of these models couldn't use the context of the original tweet and recent references. Like, say a tweet referencing the inhuman treatment of protesters in Iran for the Hijab Controversy was treated as toxic by the toxic-bert classifier due to the use of abusive words, but it only represented their conditions.

Even with the availability of contexts, the human annotators working independently weren't able to reach a consensus. This was usually found due to differences in personal opinions. For a sentence, "Stop filling false cases. Feminism is cancer...", some people wholeheartedly agreed with it; some said this was toxic as it was saying that feminism as a whole is cancer and thus is toxic towards feminists.

On the topic of, say, Asian Hate Crimes, a common observation was that most tweets the model disagreed upon talked about how Blacks were hugely responsible for Asian Hate Crimes. The 2019

Bureau of Justice Statistics, US report shows that 27.5 per cent of violent criminals targeting an Asian victim are black (12.1% of the population in the US), and only 24.1 per cent are white (57.8% of the population in the US). Some argue that these tweets were just stating the truth, and some argue they targeted a community. This was the case when the respondents were Asians, so bias is inherent.

3 Conclusion

Overall, this assignment was a great experience with lots of things to learn. The use of Machine Learning Models was a new experience for all of us. From, what we observed, Google Perspective API was quite good in the classification process except when it required context, which it didn't have. But, toxic-bert seemed to just look for words which it termed toxic and made the classification accordingly, which some nuances, of course. But, the human annotation process was more troubling, as annotators disagreed on most things. But, I believe a more objective guideline for the annotation process may help us to have a better agreement in future.