Bias, Censorship, and Accuracy: AI Models for Online Moderation

Josef Karpinski (jjk21004)
Alexander Manos (agm21019)
Samuel Cultrera (sac19019)
Malcolm Ferguson (mrf20004)

Abstract

This report presents two AI models that can be used to classify online comments as toxic or non-toxic. For this task, we used the Jigsaw Toxic Comment Classification dataset, training our own logistic regression model, as well as testing a pre-trained LLM for this task. We compared each model's ability to detect toxic or non-toxic comments, bringing up further discussion on the practical use of these algorithms and the ethical concerns behind those use cases. For results, we achieved 92% precision, 62% recall, and 95% accuracy for the logistic regression model, as well as 80% precision, 79% recall, and 96% accuracy for the LLM. As this is a course on ethics in computer science, we made sure to heavily focus on the ethical concerns behind implementing these algorithms in the real world and how it can ultimately affect end users.

Our complete project can be found here: GitHub

Introduction

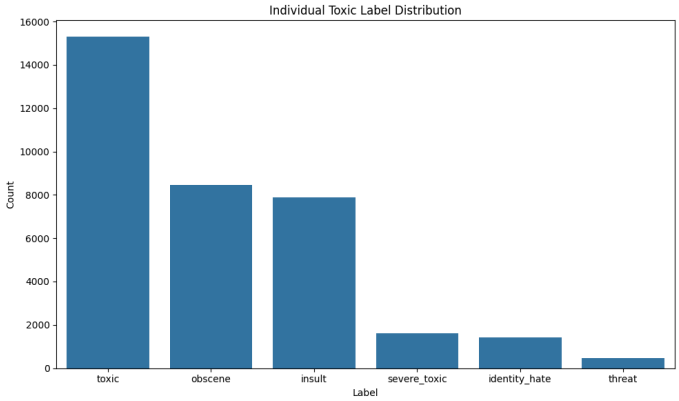
In the world of social media, there is a great emphasis we, as a society, place on monitoring online behavior to ensure the safety of everyone involved. It is well known that there are many individuals who use the ability to speak on the internet to spread hateful messages and ideas. Many social media websites and online platforms use content monitoring systems to try and limit the amount of toxic content on their platforms. This results in many questions arising about both how the content should be filtered, as well as many discussions about the ethics of filtering/censoring content, which is something that many people consider to be a form of free speech. In this project, we aim to explore both the methods of eliminating toxic content by creating an AI model to detect toxic content, as well as examine the ethical implications of censoring content.

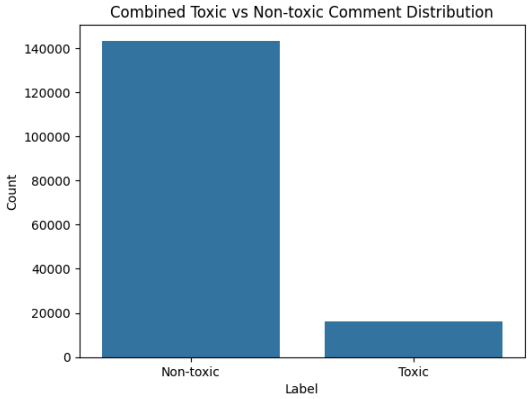
To implement the model, we selected a dataset, the Jigsaw Toxic Comment Dataset, to build the model off of. We used logistic regression to create our own model, and our goal is to use this to create a model that can accurately classify input text segments as toxic or non-toxic. We will evaluate the performance of our model using accuracy, precision, recall, and f1-score. On top of this, we compare our results to a pre-trained large language model for further analysis.

Methods

Since we will be mainly focusing on ethical concerns for this project, we will only use two different models for this project. First, we trained a simple logistic regression model, and then we will use an LLM that was pretrained on the dataset.

First, it is important to deeply understand our dataset before running methods and analyzing results. This dataset includes approximately 160,000 samples, each with a text input, and a binary label for each of the following 6 categories: toxic, severe_toxic, obscene, threat, insult, and identity_hate. For the sake of simplicity, we will combine all of these labels into a singular toxic label defined by whether any of the 6 toxicity labels are true. Below are two graphs that show some basic statistics on these labels.





It is important to note for analyzing results that about 90% of the samples in this dataset are classified as non-toxic, meaning that simply guessing non-toxic for all samples will already give us an accuracy of 90%. For that reason, it is important to not only measure accuracy, but also metrics like recall, precision, and f1-score.

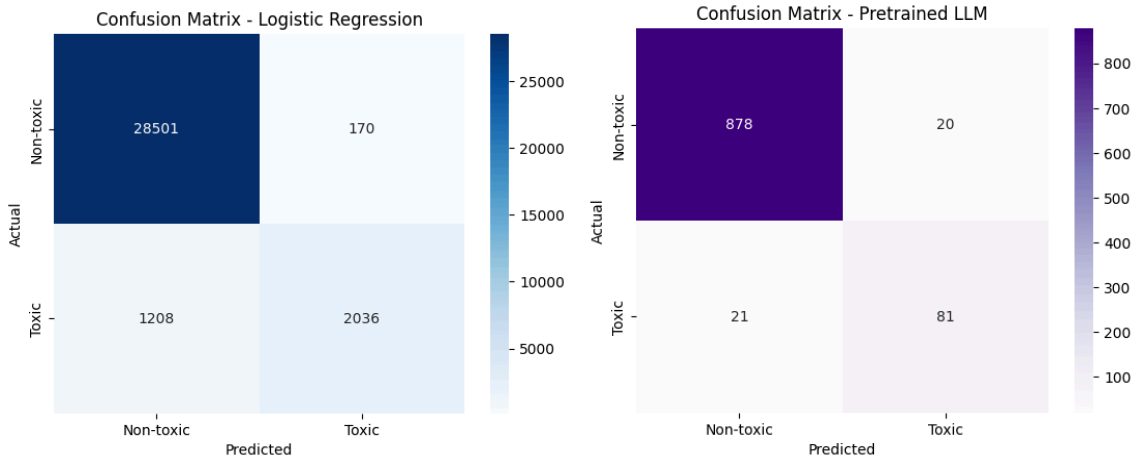
For our first model, we trained our own binary classifier using Logistic Regression, utilizing the entire dataset. We made sure to split our data into a training set and testing set using an 80-20 split. Also, before feeding our model the data for training and testing, we made sure to vectorize the text data using TfidfVectorizer.

For our second model, we used a pre-trained LLM from HuggingFace, named unbiased-toxic-roberta. It can be found here. This classifier is trained on a few different datasets, including the one we are working on in this project. The LLM outputs scores for 17 different labels, 6-7 of which correspond directly with the labels for our specific dataset. For each of these labels, the LLM outputs a score along with the label. Our general methodology will be to check each of these labels' scores, and if just one of them is beyond a certain threshold, then we will classify that text as toxic.

Results

After creating the models, we ran them against the test set and collected the results. The precision, recall, and F1-score of the toxic comments were 0.9229, 0.6276, and 0.7472, respectively. Taking into consideration that there was an imbalance with more non-toxic comments than toxic comments in the data, an f1-score of 0.8020 on the toxic comments means that the model is able to strongly identify toxic comments. The precision, recall, and f1-score of the non-toxic comments were 0.9593, 0.9941, 0.9764, respectively. These numbers are also strong, but can be more difficult due to the data containing more non-toxic comments than toxic. However, these numbers are all on the stronger end and demonstrate that the model is performing well, which explains the total accuracy of 0.9568.

Similarly, the LLM produced precision, recall, and f1-scores on toxic comments of 0.8020, 0.7941, and 0.7890, respectively. On non-toxic comments, the scores were 0.9766, 0.9777, 0.9772 respectively. The total accuracy of the LLM was 0.9590. These are slightly better than the results for the logistic regression model. One significant difference between the two is the recall score for the toxic comments, the 17% jump shows that our LLM model was much better at finding the toxic comments specifically. The results can also vary depending on the threshold chosen for the classification. For instance, if we lowered the score threshold to be 0.3, it would likely increase recall – catching more toxic comments – but at the cost of precision, leading to more false positives. Further exploration into optimizing this for different moderation goals would be a valuable next step for this project.



Discussion

Due to the usage of artificial intelligence for this project, there are several ethical concerns that need to be addressed. Any task or project that incorporates artificial intelligence or large language models must be examined from an outside lens that analyzes the potential problems that can arise.

Since no artificial intelligence model can currently perform with one-hundred percent accuracy, it is necessary for the developer to decide whether or not to focus on prioritizing over classifying text as toxic or to under-classify toxic comments. The former's goal is to protect

people at a greater rate from toxic comments and hate speech, however, this limits the freedom of speech that people have and does not properly account for the context of the comment, which can provide clarifying details into whether the message is truly toxic. When determining what to do for each type of message, our group decided to prioritize flagging comments as toxic to attempt to overprotect the user. This method could result in a less positive user experience, but it better accomplishes the goals of content moderation that we set out to achieve.

Additionally, when analyzing the data, we found that toxic comments from certain groups can be more frequently flagged. Mitigating this bias is extremely difficult as it requires the artificial intelligence to be able to fully interpret the message and understand the surrounding contexts. Furthermore, when attempting to implement this in the real world, such as on a social media site or a forum, we agreed that this would be the greatest hurdle to overcome when training the model.

When implementing this in a real-world example, it is important to once again consider the difficult choice of overmoderation for triggering false positives, or a less restrictive filter to not risk imposing on free speech. The easiest way to do this would be to also send flagged messages to be graded by a human. While this is not very efficient, it is one of the only ways to ensure accuracy when grading the toxicity of messages prior to guaranteeing one hundred percent or near it, accuracy for artificial intelligence.

Conclusion

Ultimately, when answering these questions, we determined that platforms should prioritize accuracy up to a certain point, and then, when achieving the highest realistic accuracy possible, lean towards protecting the user through inclusivity. Similarly to before, we wanted to overbalance the false negatives in order to better protect particular groups of people who are most marginalized. When using and developing artificial intelligence going forward, it is ever important to protect from toxic messages and potential hate speech while also balancing the prominent need for free speech to be maintained.

References

https://github.com/josef-karpinski/content-moderation-cse3000
https://huggingface.co/unitary/unbiased-toxic-roberta
https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/data