LSTM-Based Toxicity Classification: Technical Audit

qd2046, lc4866
Qi Dong, Luoyao Chen

May 11, 2023

1 Background

a. What is the purpose of this ADS? What are its stated goals?

The Kaggle Competition, Jigsaw Unintended Bias in Toxicity Classification [1], is a competition to detect toxic comments while minimizing unintended bias. Toxicity is defined as anything rude, disrespectful, or otherwise likely to make someone leave a discussion. In their last year's competition to classify toxic comments, they found that the models incorrectly label names with frequently attacked identity with toxicity. **Therefore, Jigsaw Conversation AI team opened this competition to build an ADS that minimizes this type of unintended bias while recognizing toxicity**. The ADS [2] that we chose to audit uses LSTM and has the highest votes.

b. If the ADS has multiple goals, explain any trade-offs that these goals may introduce.

Minimizing bias will unavoidable reduce the accuracy of the ADS. It is important to find a balance between those two goals to build a fair model that has decent performance.

2 Input and Output

a. Describe the data used by this ADS. How was this data collected or selected?

The dataset used in the ADS originates from Civil Comments platform made around 2 millions of their comments available for public, and Jigsaw extended annotations from human raters for this competition. When obtaining the toxicity label, each comment was shown to up to 10 annotators, and annotators were asked to rate the toxicity to very toxic, toxic, hard to say, and not toxic. These rating were then aggregated into fractions of annotations that fell within the former two categories. To collect identity label, which we will describe more specifically in the next section, annotators were asked to rate all the attributes that were mentioned in the comment. These annotations were then aggregated into fractional values representing fraction of raters who indicated this identity. For evaluation, toxicity score and identity label that have values larger than 0.5 are labeled as positive.

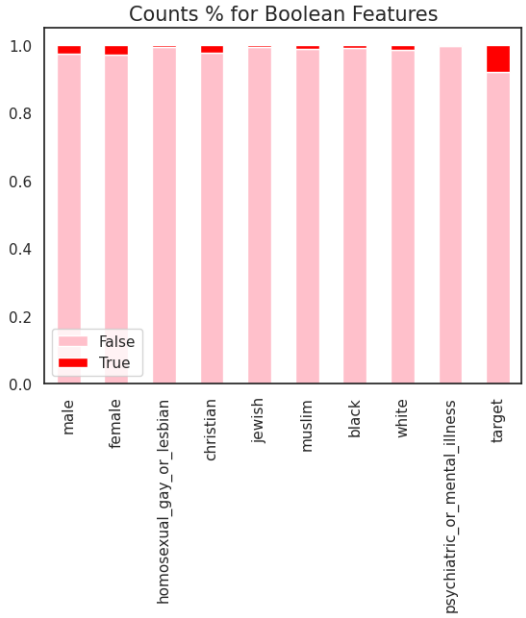
b. For each input feature, describe its datatype, give information on missing values and on the value distribution. Show pairwise correlations between features if appropriate. Run any other reasonable profiling of the input that you find interesting and appropriate.

There are 45 columns in the training data, of which 9 are "identity columns" and 6 are "auxiliary columns". In total, we used 16 features. Among them, only 1 (text) is the real input, 9 features are for bias determination, and 5 are for the output features. The datatypes are as below.

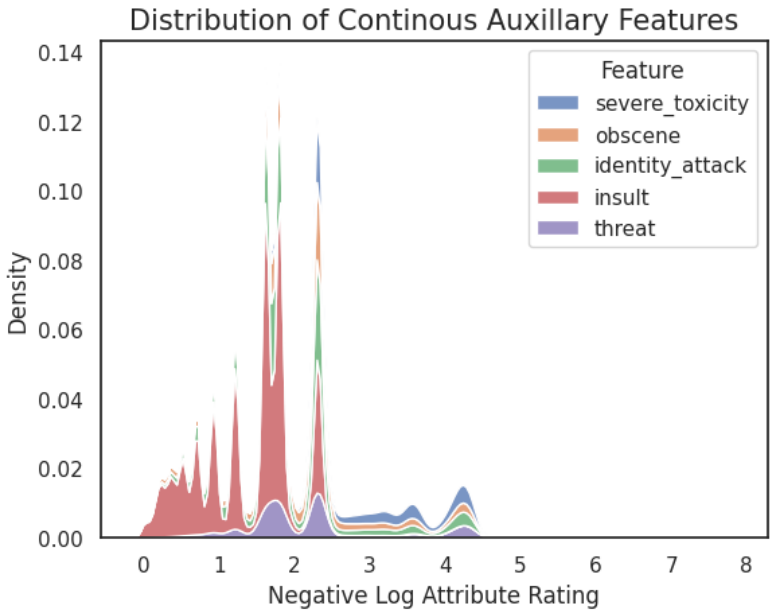
```
--------------- input feture ---------------
comment_text                        object
--------------- bias determination ---------------
male                                bool
female                              bool
homosexual_gay_or_lesbian           bool
christian                           bool
jewish                              bool
```

```
muslim                                bool
black                                 bool
white                                 bool
psychiatric_or_mental_illness         bool
--------------- output features ---------------
target                                bool
severe_toxicity                    float64
obscene                            float64
identity_attack                    float64
insult                             float64
threat                             float64
```

The count distribution of boolean features is as below. True means that the identity is present(bigger than 0.5) whereas False means that the identity is not present. The reason why most values of the identity features are False is because the identity attribute is created by labeling the comment text, which does not necessarily indicates identity attributes.



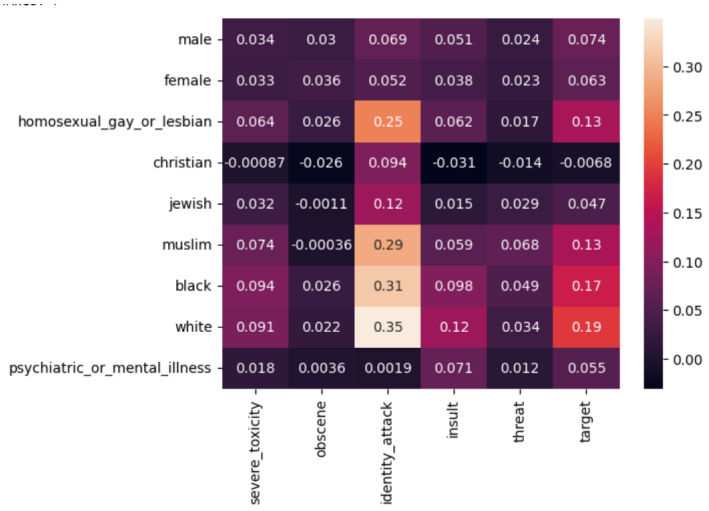
The kde plot of the auxillary features is as below (after taking the negative of the log of the ratings, to make plots more distinguishable).



From the distribution, we notice that the peak of attribute "insult" and "identity attack" are among the features that are more likely to be identified by annotators. For instance, the attribute scores of an input sentence *haha you guys are a bunch of losers* is

target	True
severe_toxicity	0.021277
obscene	0.0
identity_attack	0.021277
insult	0.87234
threat	0.0

The correlation map between the identity columns with auxiliary columns is as below:



A lot of identity attributes (input space) have a relatively large correlations with identity_attack and target (output space) attributes.

c. What is the output of the system (e.g., is it a class label, a score, a probability, or some other type of output), and how do we interpret it?

During training, there are two types of outputs that contribute to the loss:

1) Target column, which contains toxicity score, is a fractional value representing the percent of raters who believed that the comment is toxic. 2) Auxiliary features, each contains auxiliary rating for a certain input text.

During evaluation, we evaluate accuracy through binary classification accuracy i.e. the target with value larger than 0.5 is considered toxic.

3 Implementation and Validation

a. Describe data cleaning and any other pre-processing

The only data cleaning procedure is composed of feature selection, i.e. by retaining only those attributes that contain more than 500 non-null values.

To conduct data pre-processing, we used the provided training set and an expanded test set, which resulted in 1,804,874 training samples, and 97,320 testing samples. Each input text was tokenized, padded into sequence of uniform length (220). The sequences of indices were fed into the training model.

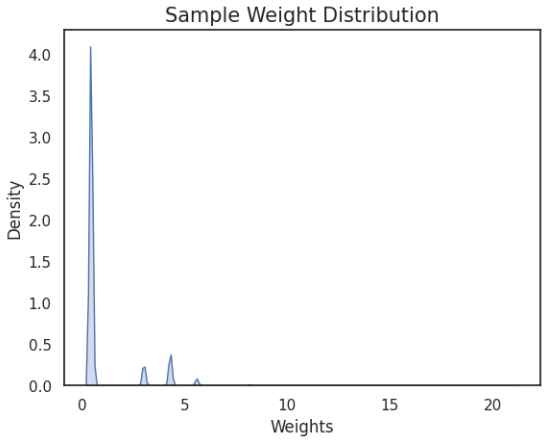
b. Give high-level information about the implementation of the system

The training structure consists of two components: model architecture, and the sample weights one use to emphasize each training sample.

In order to generate the sample weights, *identity* columns are used. Namely, higher weights are assigned to comment that either:

- Received any score on the identity columns. i.e. containing words that refer to identities (male, female, Christian. etc)
- Received a label being 1 (toxic), but does not contain any identity mentions
- Received a label being 0 (non-toxic), however, it contains identity mentions.

As a result, the distribution of sample weights on 1,804,874 training samples is as below. As can be seen from the plot, weights ranged $[0.4, 21]$, with standard deviation=1.4, mean=1. Through further investigation, all those weights larger than 5 (38049 samples) come from non-toxic comments. All toxic comments (115,355 samples) have a weight of 4.293527.



In order to train the model, an LSTM module is used, using the sample weights that assign different importance to samples (put different weights and enfore the model to focus more on the samples of interest). The ADS uses a batch size 512, trained on 4 epochs, training takes around 1.5 hr. As a result, the *aux* columns are treated as the target features, together with toxicity scores.

c. How was the ADS validated? How do we know that it meets its stated goal(s)?

The validity of the ADS was evaluated using a newly developed metric that balances overall performance with unintended bias by combining the overall metric with several sub-metrics. The overall metric used was the area under the receiver operating characteristic curve (AUC), and the ADS achieved an AUC of 0.9679 when tested on the expanded test set that was released after the competition. To assess the model's performance on different demographic subsets, various bias AUCs were calculated on each subset, each representing an identity. Subgroup AUC was based on examples that mention a specific identity, Background Positive Subgroup Negative (BPSN) AUC was based on non-toxic examples that mention the identity and toxic examples that do not, and Background Negative Subgroup

	subgroup	subgroup_size	subgroup_auc	bpsn_auc	bnsp_auc
0	black	761	0.842024	0.929058	0.937138
1	homosexual_gay_or_lesbian	538	0.842067	0.925099	0.937546
2	muslim	1054	0.846241	0.946483	0.924560
3	white	1178	0.851563	0.928634	0.941786
4	jewish	411	0.907606	0.957440	0.935474
5	male	2112	0.923295	0.964468	0.939268
6	female	2602	0.936724	0.974402	0.934392
7	psychiatric_or_mental_illness	238	0.936997	0.951018	0.960423
8	christian	2109	0.940380	0.977608	0.926534

Figure 1: Bias AUCs

Positive (BNSP) AUC was based on toxic examples that mention the identity with non-toxic examples that do not. The bias metrics were combined using power mean, and the final metric was calculated by taking the average of the overall metric and the bias metrics. Having a high final metric (0.9350) means that the ADS achieves high accuracy while minimizing the unintended bias.

4 Outcome

a. Analyze the accuracy of the ADS by comparing its performance across different subpopulations, with respect to different accuracy metrics. Carefully justify your choice of accuracy metrics.

As discussed in part c of Implementation and Validation, the competition and the ADS were evaluated using several accuracy metrics, including Subgroup AUC, BPSN AUC, and BNSP AUC. Therefore, we first calculated those accuracy metrics and the result are presented below (Fig.1).

Among the 9 identity subgroups, the black subgroup had the lowest Subgroup AUC, while the Christian subgroup had the highest Subgroup AUC. However, all subgroups achieved high BPSN AUC and BNSP AUC scores, indicating that the ADS was able to accurately distinguish between toxic and non-toxic examples that mention the identity from those that do not.

In addition to the bias metrics used in the competition, we also evaluated the ADS using other accuracy metrics that were used in lab and homework 1. We grouped the nine identity subgroups into five different sensitive groups: gender, race, religion, sexuality, and disability. However, since some of the sensitive groups, such as disability and sexuality, only had one identity subgroup, we added an additional feature called "other" to represent examples that did not belong to other identity groups or could not be recognized by the annotators. The gender feature was divided into subgroups of female, male, and other(Fig.2); the race feature into subgroups of black, white, and other(Fig.2); the religion feature into subgroups of Christian, Jewish, and Muslim(Fig.3); the sexuality feature into subgroups of homosexual and other(Fig.4); and the disability feature into subgroups of psychiatric or mental illness and other(Fig.5).

gender	accuracy	precision	recall	FNR	FPR
female	0.920364	0.817568	0.517094	0.482906	0.017717
male	0.900568	0.766990	0.493750	0.506250	0.026786
other	0.945982	0.620793	0.773778	0.226222	0.039593

(a) Gender

race	accuracy	precision	recall	FNR	FPR
black	0.805383	0.785714	0.464789	0.535211	0.052786
white	0.792869	0.773869	0.436261	0.563739	0.054545
other	0.947104	0.622438	0.775611	0.224389	0.038766

(b) Race

Figure 2: Gender and Race accuracy

The attributes from the identity column have a higher precision but a lower recall compared to the "other" column. This means that if the text is identified as belonging to one of the subgroups, it is more likely to be truly classified as

religion	accuracy	precision	recall	FNR	FPR
christian	0.945813	0.800000	0.470588	0.529412	0.010753
jewish	0.886154	0.777778	0.403846	0.596154	0.021978
muslim	0.834915	0.775701	0.356223	0.643777	0.029233
other	0.945938	0.623234	0.775583	0.224417	0.039654

Figure 3: Religion Accuracy

sexuality	accuracy	precision	recall	FNR	FPR
homosexual_gay_or_lesbian	0.795539	0.750000	0.428571	0.571429	0.057292
other	0.945362	0.625957	0.761118	0.238882	0.038886

Figure 4: Sexuality Accuracy

positive, but also more likely to be falsely classified as negative. The false negative rate is also lower as a result of this imbalance.

b. Analyze the fairness of the ADS, with respect to different fairness metrics. Carefully justify your choice of fairness metrics.

We applied similar methods to the accuracy metrics and divided the subgroups into five groups, and analyzed their fairness using metrics from lab and homework 1, such as false negative rate difference, false positive rate difference, demographic parity ratio, equalized odds ratio, and selection rate difference.

	fnr_difference	fpr_difference	demographic_parity_ratio	equalized_odds_ratio	selection_rate_difference
gender	0.280028	0.021877	0.863117	0.447464	0.013351
race	0.339350	0.015780	0.545432	0.562473	0.079055
religion	0.419359	0.028902	0.485245	0.271161	0.052257
sexuality	0.332546	0.018406	0.585515	0.563082	0.067797
disability	0.051184	0.004523	0.496223	0.895970	0.097369

Features in the identity column tend to have a higher false negative rate than "other" features. Therefore, the false negative rate difference (between the highest and lowest subgroup) is relatively high for all groups except disability, and the equalized odds ratio is mostly lower than 0.6 since both metrics are associated with the number of false negatives. Additionally, the demographic parity ratio is lower than 0.6 for all groups except gender, indicating imbalanced selection rates across different groups.

c. Develop additional methods to analyze ADS performance: think about stability, robustness, performance on difficult or otherwise important examples (in the style of LIME or SHAP), or any other property that you believe is important to check for this ADS. Carefully justify your methodology.

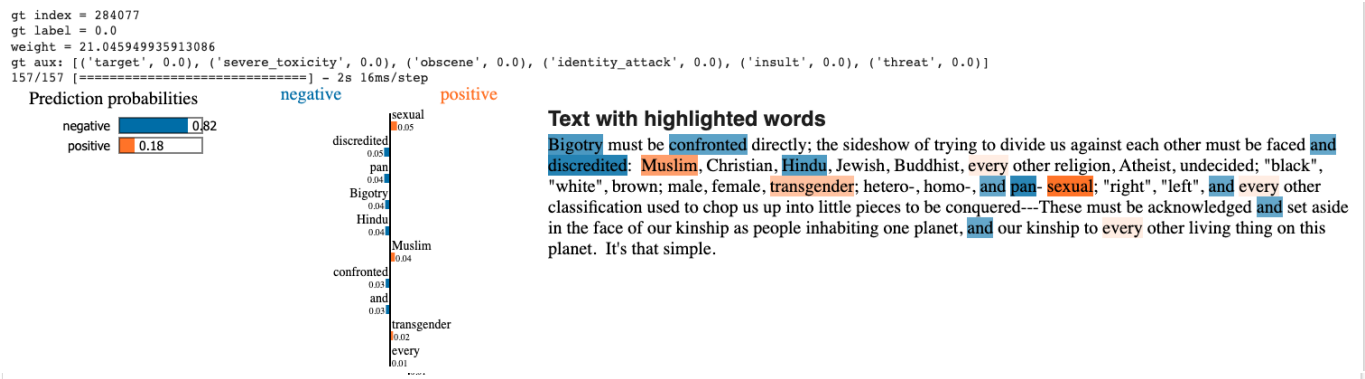
Using LIME [3], we approached the analysis from two perspectives. 1) To investigate stability, we investigated the training samples that were important during training, i.e. indicated by samples having high weights) 2) To test

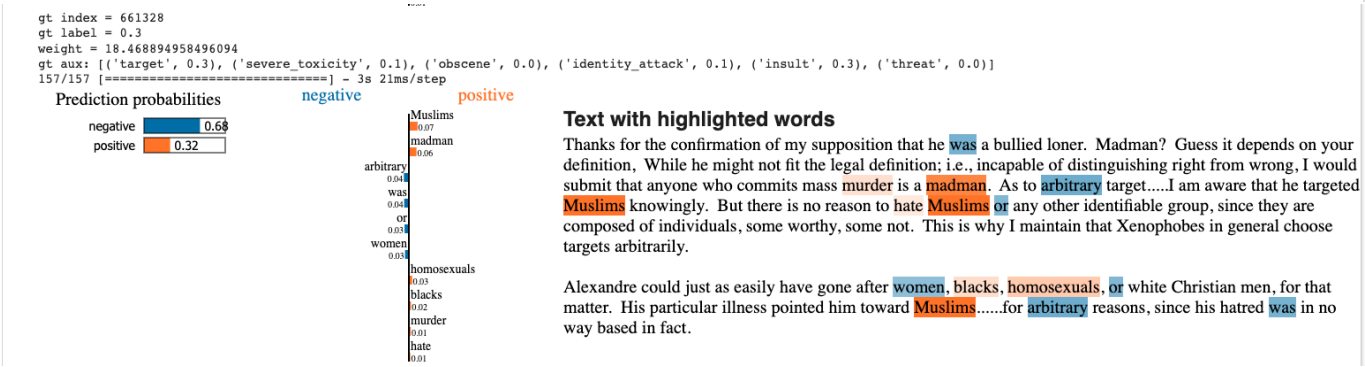
disability	accuracy	precision	recall	FNR	FPR
psychiatric_or_mental_illness	0.899160	0.826087	0.703704	0.296296	0.043478
other	0.944645	0.626141	0.754888	0.245112	0.038955

Figure 5: Disability Accuracy

robustness, we investigated the ADS's performance on the out-of-distribution (OOD) samples (expanded dataset), especially, we highlighted three FN, FP samples each.

To investigate the training samples, we investigated 2 samples of the highest weights and 2 with the lowest weights. Two samples of the highest weight (21.046 18.469) are:





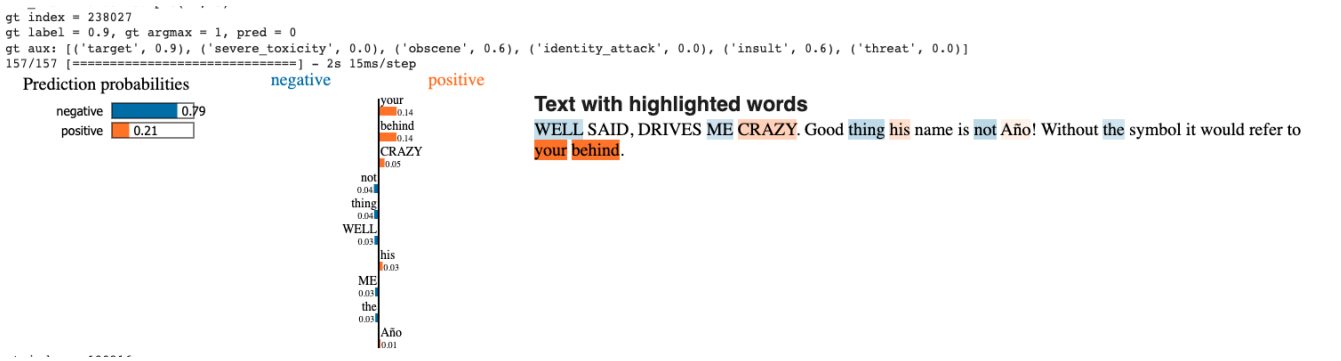
Two samples of lowest weights (0.429, 0.429) are:

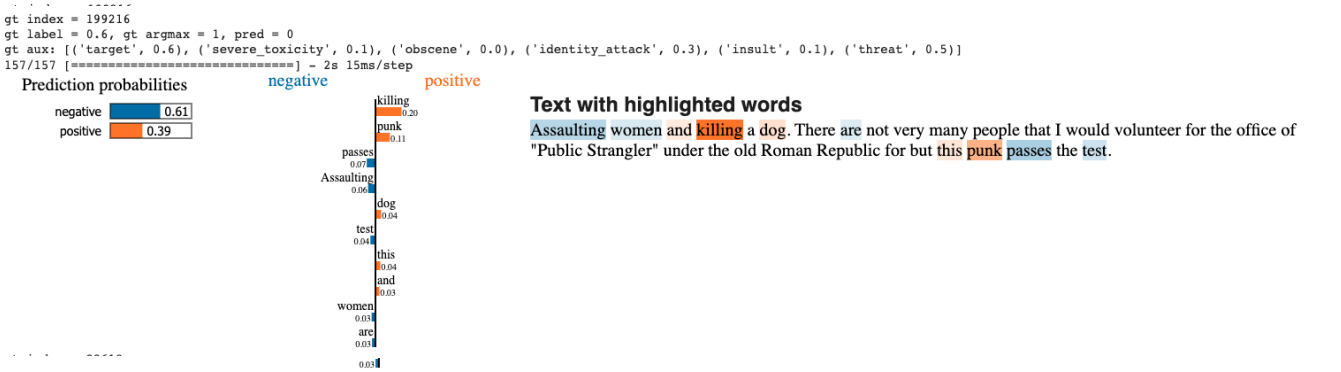
Through comparisons within the two sets of training samples, we concluded that longer sentences receive higher weights as they are more likely to contain the identity features. Moreover, words such as **Muslim**, **black** appear frequently when contributing to positive (toxic) classifications.

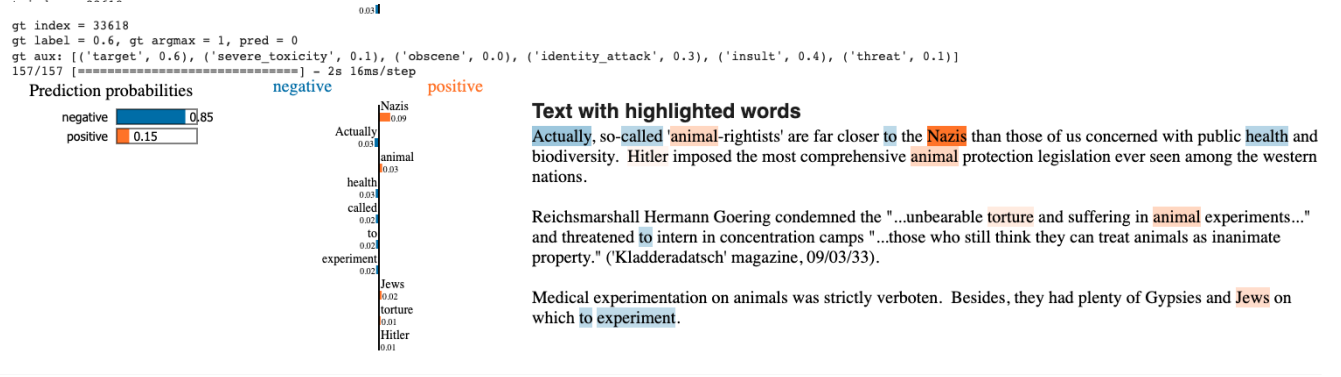
To investigate the performance on OOD samples, we measured the overall performance on the expanded set (97,372 samples). Noticeably, the classifier received a relatively high F1 score (68%).

```
Accuracy of the model :  0.9445334977394163
F1-score:  0.6849538928446364
Confusion matrix:
        [[86054  1909]
         [ 3489  5868]]
```

To further investigate, we looked into several FN, FP instances. Three samples of FN instances are:







Interesting conclusions can be drawn from OOD samples. For FN samples (ground truth being 1 while pred being 0), we conclude that the ground truth being toxic (strong, toxic sentiment) could be indicated by either word capitalization (see example.1, where the user uses "WELL SAID" to convey anger and indicates toxicity), or by only one word (see word "punk" in example.2), as well as the fact that most of the FN ground truth labels are close to neutral (toxicity closer to 0.5 in all three examples). On the other hand, however, the LSTM classifier misses those signals due to it being trained using all lowercase tokenizations, and the prediction on the entire sentence will be misguided by certain words (other than "punk" in example.2), or the prediction score falls narrowly lower than 0.5. As a consequence, all those examples are mis-classified as non-toxic (all three examples).

Three samples of FP instances are as below.

For FP samples (i.e.ground truth being 0 while pred being 1), we conclude the misclassification happens due to some misleading, biased words. For instance, see example 3, where the short sentence is neutral, however, within 10 words, it contains 2 identity words (American, Islamist), which would relate to it being classified as toxic.

5 Summary

a. Do you believe that the data was appropriate for this ADS?

Data quality can be justified from two aspects, data profiling as well as the content of comments. As for class distributions, there are apparent imbalances in this dataset, indicated by the overly-dominating distribution of the non-toxic class (99%). This will result in the debiasing being less effective (the highest weighted samples are those with label non-toxic). The content of the text is mainly about politics and public figures and hence is suitable for the sake of toxicity classifications.

b. Do you believe the implementation is robust, accurate, and fair? Discuss your choice of accuracy and fairness measures, and explain which stakeholders may find these measures appropriate.

The implementation of the ADS is accurate and robust, with a relatively high overall accuracy and AUC, and it generalizes well to unseen test data. This makes the ADS highly beneficial for Jigsaw, the host of this competition. When calculating the bias metrics used to evaluate the implementation in the competition, such as Subgroup AUC, BPSN AUC, and BNSP AUC, the ADS performs relatively well on all subgroups. However, when calculating the general accuracy metrics that were used more often, we found that text identified as belonging to one of the subgroups

tends to have a higher precision and false negative rate, and lower recall. As a result, some fairness metrics, such as false negative rate difference, demographic parity ratio, and equalized odds ratio, do not perform well. Individuals from sensitive subgroups who do not leave a toxic comment benefit from the ADS because they are not falsely classified as toxic. Additionally, individuals from sensitive subgroups who do leave a toxic comment may benefit as they could potentially be falsely classified as non-toxic.

c. Would you be comfortable deploying this ADS in the public sector, or in the industry? Why so or why not?

One possible reason why the ADS performs worse when evaluating using common accuracy metrics compared to the bias metrics used in the competition is that AUC calculations involve probabilities, while general accuracy metrics involve binary predictions where values greater than 0.5 are labeled as positive. If the public sector uses probabilities to make decisions, then deploying this ADS would be acceptable. However, if the public sector uses binary values of either 0 or 1 to make decisions, I would not be comfortable doing so. In the latter, adjustments might be necessary to achieve better fairness metrics.

d. What improvements do you recommend to the data collection, processing, or analysis methodology?

Improvements come from several aspects, corresponding to the mis-classifications.
1) For the data collection and processing methodology, more procedures are needed to address the class imbalance issue.
2) Capital/lower case needs to be taken into account during tokenization, since the same word, once capitalized, can convey toxic sentiments.
3) The weight of each word in the input should be adjusted so that it embeds prior knowledge (for instance, the word "punk" conveys strong toxicity, and hence the presence of word "punk" should have higher impact in the classification).
4) The toxicity threshold of 0.5 needs more tuning. As most comments has a lower toxicity, they all fall within the non-toxic category. As a result, it requires a granular threshold for those comments whose toxicity falls lower than 0.5.
5) The logic of the weight adjustments needs more adjustments, instead of adding weights to three scenarios, focusing on the effect of one set of features might be more persuasive.

References

[1] Jigsaw, "Jigsaw unintended bias in toxicity classification," 2019. Last accessed 16 March 2023.

[2] Thousandvoices, "Simple lstm," Jun 2019.

[3] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.