DSGA-1017
Spring 2022
Chloe Zheng, Preston Harry

Project Draft (due 4/15)

(1) Background: general information about your chosen ADS

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

The purpose of this ADS is to predict the "toxicity" of comments made on the Civil Comments platform–an application to allow users to comment on and discuss articles from independent news sites while regulating the toxicity of those comments. "Toxic" comments are defined as anything *"*rude, disrespectful or otherwise likely to make someone leave a discussion." The ADS should flag toxic comments to regulate online conversations and ideally protect voices that are discouraged from participating online. Past models have had difficulty with predicting toxicity pertaining to one's identity or other protected characteristics by associating any mention of the protected characteristic with toxicity. Any ADS should both identify toxic comments while mitigating harmful bias related to user's identities.

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

However, simply maximizing accuracy may exacerbate these biases by disproportionately assigning toxic labels to certain groups, causing disparate treatment in comment regulation. Meanwhile, guaranteeing an equitable distribution of toxicity classification may reduce prediction accuracy. The challenge will be to reasonably balance fairness and accuracy such that the ADS that does not disparately impact those with protected characteristics while still retaining a high enough accuracy to be useful.

(2) Input and output

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

This ADS uses data containing all public comments and associated commenter data from the Civil Comments application from its opening in 2015 to its closing in 2017. These comments total to about 1.8 million. Jigsaw, a Google subsidiary focused on understanding technology's role in social issues, used annotators to manually label comments as severe, obscene, an identity attack, insulting, threatening and/or sexually explicit. The final dataset includes a target indicating whether a comment is toxic, the text of the comment, labels to describe the type of toxicity, and a series of comment topic indicators including gender, sexuality, religion, and race.

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.

The input of the model only has one feature “comment text,” where each observation represents one comment from a given user.

Each comment is converted to a vector of tokens with each token being represented by an integer value in the order those tokens appear in the comment. Comment vectors are zero padded at the beginning of each input sequence such that every vector is of length 250. Therefore, a comment consisting of *n* tokens is represented by a one-dimensional vector of length 250 where the last *n* values of the vector are the tokens of the comment in order. Randomly sampling 10% of the dataset, we found no missing or empty comments in the sample. Since there is only one feature being inputted in the model, it is not relevant to calculate pairwise correlations of the input data.

The full dataset includes several features that indicate how a comment was rated as well as whether the comment’s content pertains to particular protected characteristics. Table 1 presents for each feature, the type, proportion of missing values, and descriptive statistics for numeric features. Certain subjects appear very infrequently in the test data and the features indicating those subjects have missing features in the training data. These features, listed red in Table 1, have the same proportion of missing values and require further investigation to understand why.

Some additional features describe the type of toxicity of a comment according to human graders. The most frequent type of toxicity are insults while “severely toxic” comments are least frequent. The pairwise correlations are displayed in Figure 1. Severely toxic, obscene, identity attack, and insult identifiers are all somewhat correlated with one another, suggesting that these descriptors are associated with one another in the eyes of the graders. Threats and sexually explicit comments have less association with the other types of toxicity.

Table 1:

feature	**type**	**missing_prop**	**mean**	**med**	**std**	**min**	**max**
comment_text	object	0	-	-	-	-	-
severe_toxicity	float	0	0.004656	0	0.022997	0	0.6
obscene	float	0	0.014258	0	0.066188	0	1
identity_attack	float	0	0.022551	0	0.078741	0	1

insult	float	0	0.081502	0	0.176302	0	1
threat	float	0	0.009279	0	0.049485	0	1
asian	float	0.775989	0.011933	0	0.086801	0	1
atheist	float	0.775989	0.003121	0	0.049543	0	1
bisexual	float	0.775989	0.001926	0	0.027667	0	1
black	bool	0	0.008549	-	-	-	-
buddhist	float	0.775989	0.001635	0	0.032864	0	1
christian	bool	0	0.022517	-	-	-	-
female	bool	0	0.02972	-	-	-	-
heterosexual	float	0.775989	0.003238	0	0.045688	0	1
hindu	float	0.775989	0.001659	0	0.03268	0	1
homosexual_gay_or_lesbian	bool	0	0.006045	-	-	-	-
intellectual_or_learning_disability	float	0.775989	0.001075	0	0.016295	0	0.8
jewish	bool	0	0.004022	-	-	-	-
latino	float	0.775989	0.005805	0	0.056513	0	1
male	bool	0	0.024894	-	-	-	-
muslim	bool	0	0.011735	-	-	-	-
other_disability	float	0.775989	0.001168	0	0.013535	0	0.4
other_gender	float	0.775989	0.00086	0	0.011461	0	0.25
other_race_or_ethnicity	float	0.775989	0.008118	0	0.041905	0	0.857143
other_religion	float	0.775989	0.006666	0	0.037724	0	1
other_sexual_orientation	float	0.775989	0.001457	0	0.014917	0	0.5
physical_disability	float	0.775989	0.00127	0	0.016765	0	0.6

psychiatric_or_mental_illness	bool	0	0.002804	-	-	-	-
transgender	float	0.775989	0.006833	0	0.069999	0	1
white	bool	0	0.013951	-	-	-	-
created_date	object	0	-	-	-	-	-
rating	object	0	-	-	-	-	-
funny	int	0	0.278258	0	1.034031	0	52
wow	int	0	0.044125	0	0.244807	0	11
sad	int	0	0.109	0	0.450457	0	14
likes	int	0	2.432962	1	4.664847	0	201
disagree	int	0	0.583815	0	1.8777	0	107
sexual_explicit	float	0	0.006651	0	0.045983	0	1
identity_annotator_count	int	0	1.46446	0	19.34003	0	1854
toxicity_annotator_count	int	0	8.753705	4	42.17269	3	3623

Figure 1:

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?

For a given comment, the output of the model is the probability that the comment is toxic. Given some threshold, if the probability is above the threshold the comment is toxic, otherwise it is not. The developers do not indicate a threshold, so assume the threshold is 0.5.

(3) Implementation and validation

Present your understanding of the code that implements the ADS. This code was implemented by others (e.g., as part of the Kaggle competition), not by you as part of this assignment. Your goal here is to demonstrate that you understand the implementation at a high level.

a. Describe data cleaning and any other pre-processing

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

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

Plan:

1. *Run and explain "Load and preprocess the data set" section in provided notebook*
    a. *Describe preprocessing of text data*
    b. *Many descriptive features are missing at the same rate–investigate and understand why these features have missing values: do only the same observations have missing values? What kinds of comments are they missing for?*
2. *Run the convolutional neural network on a sample of the dataset.*
    a. *Describe embeddings process*
    b. *Describe layers, activation functions, etc.*
    c. *Describe hyperparameters used*
3. *Run model on validation dataset*
    a. *Describe how much data is used in validation set*
    b. *Describe how they evaluate validation set predictions - run their AUC bias metrics*

(4) Outcomes

a. Analyze the effectiveness (accuracy) of the ADS by comparing its performance across different subpopulations.

b. Select one or several fairness or diversity measures, justify your choice of these measures for the ADS in question, and quantify the fairness or diversity of this ADS.

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

Plan:

4. *Discuss existing subpopulation AUC breakdown*
5. *Expand existing AUC discussion*

    a. *Look at additional subgroups–others exist in the data but have many missing values–additionally it may be helpful to look at "missing" data as its own subgroup*
    b. *Look into FPR and FNR by subgroup*
6. *Determine and justify fairness measure*
7. *Attempt to apply feature analysis (LIME, SHAP)*
    a. *The neural net is built on one input feature, so some of these methods probably won't work well at first–figure out some way to implement them in this context if possible*

(5) Summary

a. Do you believe that the data was appropriate for this ADS?
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.
c. Would you be comfortable deploying this ADS in the public sector, or in the industry? Why so or why not?
d. What improvements do you recommend to the data collection, processing, or analysis methodology?

Plan:

8. *Determine in what sector this ADS would be deployed*
9. *Most of the summary will depend on the outcomes of previous sections and will be written afterwards*
    a. *Some parts can be written in advance without extra analysis (part a and some of part d)*