Nutritional Label for Automated Decision Systems

-

Toxicity Classification

Sarath Kareti

May 9, 2022

Background

Purpose

This ADS detects toxicity in online comments while simultaneously attempting to minimize unintended model bias. Toxicity is defined in this case as any comment that is “rude, disrespectful or otherwise likely to make someone leave a discussion”. This ADS will attempt to detect toxicity in comments without automatically exhibiting bias and assuming that all comments containing words such as frequently attacked identities (i.e. “black”) are toxic. The ADS will attempt to specifically denote comments as toxic when they are actively attacking the minorities, not just mentioning them, and thereby minimizing bias against these minority groups in the mode.

Tradeoffs

As mentioned before, the first goal of this ADS is to identify comments in online conversations that are toxic. The second goal of this ADS is to ensure that this occurs without exhibiting bias, and to differentiate between comments that are attacking protected minority groups and those just mentioning them. What most likely will have to be sacrificed in exchange for achieving these goals is accuracy of the model overall. The two choices with this ADS will result in either an increase in the number of false positives of the ADS, or an increase in the number of false negatives of the ADS. In the case that there is no attempt to minimize the bias exhibited by the system, the number of false positives will increase, as many comments will be flagged for toxic (false positive result for toxicity) even though the comment was not toxic. In the case that there is an attempt to minimize the bias exhibited by the system, the number of false negatives will increase, as many comments that are toxic may be missed by the system. In either case, accuracy will decrease.

Background

Data selection/collection

The data that is used by this ADS has been collected from the Kaggle competition “Jigsaw Unintended Bias in Toxicity Classification”. This competition offered $65,000 of prize money to the winner. The data was collected by Kaggle from the Civil Comments platform, after the platform shut down and made available its archive of roughly 2 million comments for study and use.

Data description

The data from the original competition contains is a 1804874x45 (i rows, j columns) dataset. The “comment_text” column contains the original text of the comment. The “target” column contains the toxicity score of the comment. There are columns that detail toxicity subtypes such as “severe_toxicity”, “obscene”, “threat”, “insult”, “identity_attack”, and “sexual_explicit”. There are also columns that contains whether or not certain identities are mentioned in a comment, such as “male”, “female”, “transgender”, “other_gender”, “heterosexual”, “homosexual_gay_or_lesbian”, “bisexual”, “other_sexual_orientation”, “christian”, “jewish”, “muslim”, “hindu”, etc. There are also columns that contain miscellaneous information about columns such as "created_date", "parent_id", and "identity_annotator_count".

The “comment_text” column contains a string. The “id” column contains an int. The “target” column contains a floating point number that is between 0 and 1 and denotes the toxicity of the comment. The other toxicity and identity labels are also floating point numbers between 0 and 1.

The missing values in this dataset are mostly found in the identity columns, as if the comments don't contain certain identities, those columns for those identities for those comments will be empty. Just over 20% of the dataset contains comments that are labeled with identities. The rest of the dataset contains missing values for the identity columns.

Pairwise Correlation

After studying the relationships between the different variables and columns in the dataset, the highest mutual information score was found between the “target” column and the “insult” column, with a score of 0.65. This means that there is a significant correlation between whether comments are considered toxic or not and whether they are classified as insults. The next highest mutual information score seems to be found between the columns “severe_toxicity” and “obscene” with a score of 0.39. Following that is the score between “target” and “identity_attack” which was found to be 0.33. Also, further analysis made it very clear that when a comment mentions one of the identities "homosexual_gay_or_lesbian", "black", "white", or “muslim”, it is more likely to be classified as toxic.

Output

This ADS is a binary classification model, and the output is a representation of whether or not the comment is toxic. Though binary indicates that the output would be either a 0 or a 1, the system rather outputs a number between 0 and 1 that represents the probability that a comment is toxic or not. If the output is closer to 1, the comment is more likely to be toxic. If the output is closer to 0, the comment is more likely to not be toxic. The threshold that this system will be using to decide whether a comment is toxic or not is an output of 0.5. Any comment with a score under 0.5 will be considered not toxic, and any comment with a score over 0.5 will be considered toxic.

Implementation and Validation

Data cleaning and pre-processing

The training data was read into a dataframe called `train_df`. The validation or testing data was read into a dataframe called `test_df.` The data was then cleaned with a TFID vectorizer. This was done to convert the text into feature variables which are easily used as inputs for the model. During cleaning, punctuation was removed from the comments to set up for smoother analysis.

High level overview of implementation

After the pre-processing and data cleaning steps were completed, a Naive-Bayes-Support Vector Machine classifier was created. NB-SVM models are able to run while using a much smaller amount of resources in comparison to other models which are more complicated, but the performance of these models can vary greatly based on a large variety of factors. The model was created with the following parameters: C=1.5 and dual=False set by the creator of the notebook. Predictions were made using the test set.

This model takes a logistic regression model and adds in a Naive Bayes log-count ratio matrix. The general pipeline for this ADS involved: Cleaning the comments/converting the target and identity values into booleans, tokenizing the comments and converting them to features, vectorizing the features by using a TDIF vectorizer, transforming those vectorized features into a Naive Bayes log-count ratio matrix, and feeding that matrix into the logistic regression model that was constructed.

Validation

It is important to note that the code supplied by Kaggle did not *validate* the ADs. The novel metric that was used by Kaggle to validate the solutions was a combination of different submetrics. The different submetrics were combined for the sake of balancing the tradeoff between performance/accuracy and unintended bias. This combination of submetrics was used to validate the solutions. The metrics that were used include: Overall AUC (the ROC-AUC for the full evaluation set), Bias AUCs, and Generalized Mean of Bias AUCs. Bias AUC’s included Subgroup AUC , which is where we restrict the data set to only the examples that mention the specific identity subgroup, BPSN (Background Positive, Subgroup Negative) AUC, which is where we restrict the test set to the non-toxic examples that mention the identity and the toxic examples that do not, and BNSP (Background Negative, Subgroup Positive) AUC, which is where we restrict the test set to the toxic examples that mention the identity and the non-toxic examples that do not. The Generalized Mean of Bias AUCs combine the per-identity Bias AUCs into one overall measure.

Outcomes

Analyzing accuracy

After looking at the solution and the attributes that were analyzed in the solution, we selected 3 attributes to analyze: race, sexuality, and gender. These attributes also had the highest toxicity levels out of all the attributes. Gender is separated into two categories: The privileged category, “male”, and the unprivileged category, “female”. Race is separated into two categories as well: The privileged category, “white”, and the unprivileged category, “black”. Sexuality is separated into two categories: The privileged category, “Not homosexual/gay/lesbian”, and the unprivileged category, “homosexual/gay/lesbian”. We set cutoff thresholds based on the calculated mean for each protected attribute. Any comment that contained one of the attributes mentioned was used in the analysis conducted on the attributes. The fairness metrics that we studied included: Mean Outcomes, FPR, Disparate Impact, Equal Odds, and FNR. FPR and FNR would allow us to understand the reasons behind accuracy that was lower than expected. Mean Outcomes and Disparate Impact will allow us to clearly determine which groups have advantages. Equalized Opportunity Difference allows us to further study the true positives found in any protected attribute group, where the ones that were determined by the model to be toxic were actually toxic.

Fairness/Diversity Measures

This ADS has an accuracy of 0.943 and an AUC of 0.817, which indicates that the accuracy is relatively high, and the AUC is relatively high for the model as well, though not as high as the accuracy. This indicates that the model is able to determine between positives and negatives relatively well. The precision and recall for this model were somewhat decent, with precision at 0.577 and recall being 0.671. All of these facts imply that the model performs reasonably well, but there is some room for improvement.

(i) Race

The accuracy for both the privileged and unprivileged group were fairly similar and neither were very high.

Summary	**Privileged (White)**	**Unprivileged (Black)**
● Mean Outcomes: 2.163 ● Disparate Impact: 0.709 ● EOO: 0.00114	● Accuracy: 0.759 ● False Positive Rate: 8.426 ● False Negative Rate: 93.182	● Accuracy: 0.760 ● False Positive Rate: 6.415 ● False Negative Rate: 95.876

(ii) Sex

The accuracy for both the privileged and unprivileged group were fairly similar here as well, but they were higher than the accuracies in the race subgroup.

Summary	**Privileged (Male)**	**Unprivileged (Female)**
● Mean Outcomes: -0.192 ● Disparate Impact: 0.985 ● EOO: 0.00138	● Accuracy: .8314 ● False Positive Rate: 7.935	● Accuracy: 0.840 ● False Positive Rate: 7.994

	● False Negative Rate: 93.940	● False Negative Rate: 92.661

Sexuality

The accuracy for both the privileged and unprivileged group were less similar to each other in this subgroup than the other subgroups. The accuracy in the unprivileged group is significantly lower than the accuracy of the privileged group. There was an extremely low disparate impact in this subgroup in comparison to the other subgroups, but this may result from the ADS not being able to identify this subgroup as well as the others rather than it implying a lower disparate impact in practice.

Summary	**Privileged (Not Homosexual, Gay, Lesbian)**	**Unprivileged (Homosexual, Gay, Lesbian)**
● Mean Outcomes: -2.254 ● Disparate Impact: 0.0342 ● EOO:-0.00909	● Accuracy: 0.852 ● False Positive Rate: 8.244 ● False Negative Rate: 98.2667	● Accuracy: 0.775 ● False Positive Rate: 9.820 ● False Negative Rate: 88.889

Additional Methods for Analysis of Performance

To further analyze the ADS that we are studying, we used the Shapley Additive Explanations (SHAP) to simplify and understand the output of the ADS. To explain the outcomes and output of the ADS that we are studying, we imported and used the SHAP package.

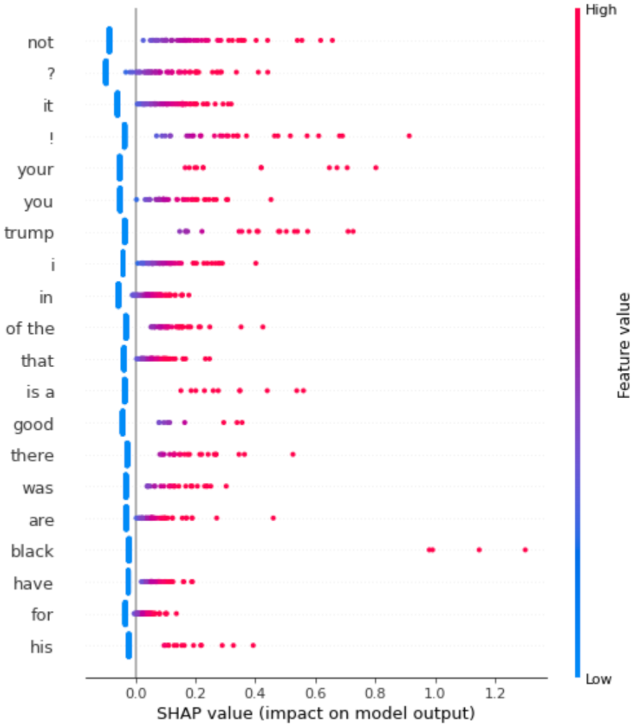
SHAP Analysis per Subgroup

I fed the SHAP explainer a smaller subset of the data so that the dataset was small enough for the explainer to work with. We randomly sampled approximately 250 entries and then conducted analysis on this subset. This allowed me to conduct analysis that was representative of as much of the true dataset as possible without overloading the google colab notebook or running out of memory.

To conduct more useful analysis, I sampled the 250 entries from each subgroup mentioned earlier so that we could determine the differences in the impacts of features on each subgroup.

Subgroup 1: Black

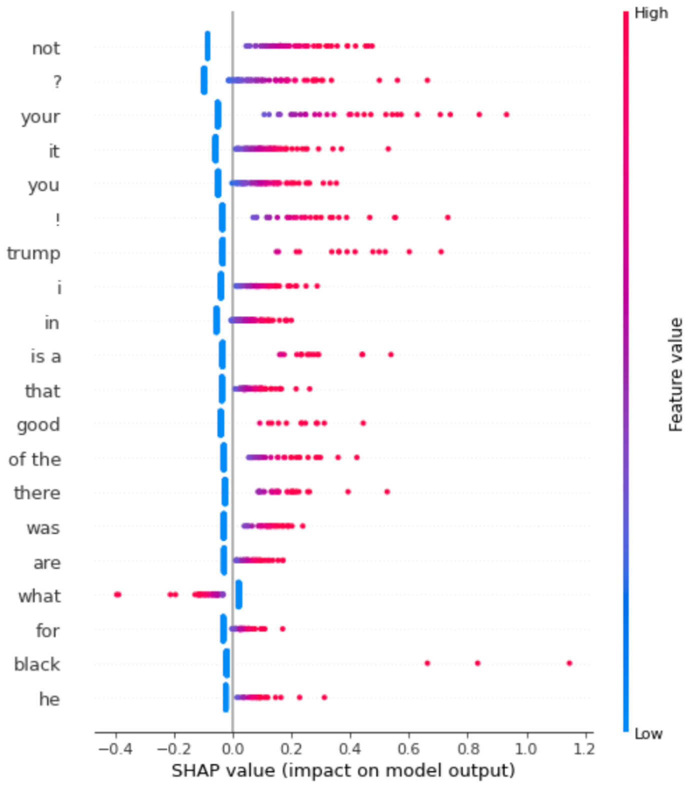
The SHAP feature that provides us with a summary plot allows us to see a ranking of the features that impact the model for the black subgroup. From this plot, we are able to determine that the top 3 impacting features that affect the model's analysis of comments within the black subgroup. It appears that the factors that impact this subgroup most include "!", "not", and "?", which makes sense as these often clearly indicate negativity in comments, leading them to be more likely to be classified as toxic.



One can also see very clearly that there are very large values for the word "black". Though it is found towards the bottom of the list, and it doesn't rank very high on the features that affect toxicity for the black subgroup, the occasions on which it does appear have very high SHAP values, indicating that the word "black" may be learned by the system to be toxic.

Subgroup 2: Homosexual, Gay, or Lesbian

Here we are shown the summary plot allows us to see a ranking of the features that impact the model for the homosexual, gay, or lesbian subgroup. Similarly to the black subgroup that was studied earlier, the identity labels "homosexual", "gay", or "lesbian" don't seem to be features that heavily affect the toxicuty of comments in the subgroup. Surprisingly, the word "black" seems to make its way into the feature list of this subgroup as well.



There are multiple reasons that this might be happening. This subgroup contains many different identity terms, including "homosexual", "gay", and "lesbian", causing them to possibly be scattered. In contrast, the word "Black" is one of the only identity terms for the African American subgroup. As a result, the model may focus on the word "Black" rather than the other words for the homosexuality subgroup such as "gay" and "lesbian".

This also may result from our analysis being unrepresentative of the whole population due to a small sampling size. This could be bad because it could lead to cognitive bias. There also could be issues in this analysis because comments that mention the homosexuality subgroup might also mention the black subgroup, causing the model to be confused by comments that mention both identities. Considering the time and resource limitations that we have been given for this project, we were not able to do the extensive analysis required to determine the overlap in identities that may cause the model to be misclassifying comments. As a result, I m not able to determine whether the model has learned that homosexuality is an indicator of toxicity or not.

General Findings

I noticed two significant things during our analysis that seemed significant during our analysis, and these things might be negatively affecting our entire dataset.

a. Small negative SHAP values

From the graphs, it is apparent that all of the features have extremely small negative SBAP values. Meaning that they have negative SHAP values that are very close to 0, and these values are not that significant. This indicates that the comments are mostly given moderately negative scores along with a small number of extremely positive values. In summary, the model is learning to use most of the features as non-toxic indicators with moderate scores, while using a few of the features as indicators with extreme scores.

b. Mis-tokenization

Sometimes, two word phrases have been classified as features rather than being separated into separate word. This has happened multiple times as shown in the graphs produced by the SHAP analysis, which showed that “of the” and “is a” were classified as features rather than being separated into the individual words. This could cause issues during the pre processing of the data and could cause issues when the individual features might affect the toxicity of some comments.

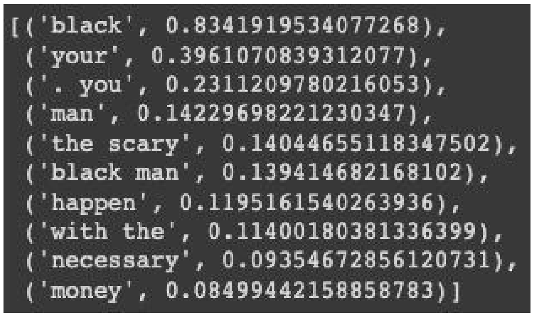
SHAP Analysis for Data Points

To further analyze the details of the ADS, we will look in detail at multiple specific points of data, and determine which features contribute to the toxicity of each comment and which features contribute to the non toxicity of each comment. During this analysis, we will be looking at comments within the “black” and “homosexual/gay/lesbian” subgroups and looking at the TP and FN calculations for both.

Subgroup 1: Black

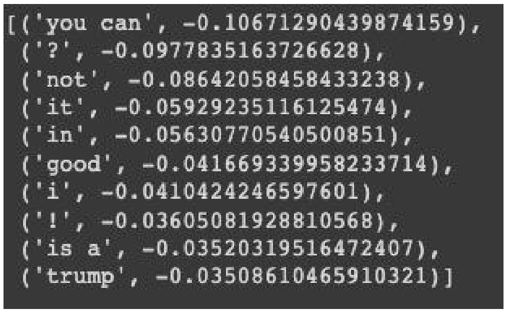
In my notebook analysis, I show the top results for positive weights and negative weights, which means that we display the top results for toxicity indicators and the top results for non-toxicity indicators. There are some results that make sense, such as seeing “black” in the toxic weights true positive category, but there are some results that donn’t make sense, such as seeing “trump” in the non-toxic weights true positive category for the black subgroup. This further proves that there may be an issue with the model’s classification. Seeing “women” in the toxic false positive category as well shows that the ADS might have also incorrectly learned this to be a toxicity indicator.

Black Toxic Weights True Positive



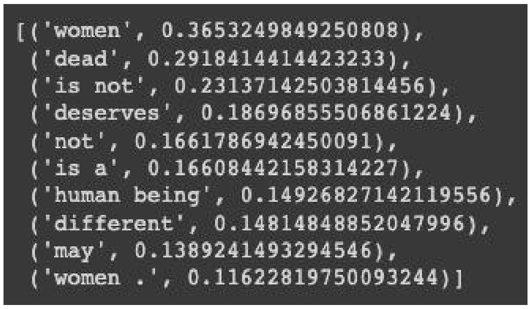
```
[('black', 0.8341919534077268),
 ('your', 0.3961070839312077),
 ('. you', 0.2311209780216053),
 ('man', 0.1422969822123O347),
 ('the scary', 0.1404465511834750 2),
 ('black man', 0.13941468216810 2),
 ('happen', 0.1195161540263936),
 ('with the', 0.11400180381336399),
 ('necessary', 0.09354672856120731),
 ('money', 0.08499442158858783)]
```

Black Non-Toxic Weights True Positive



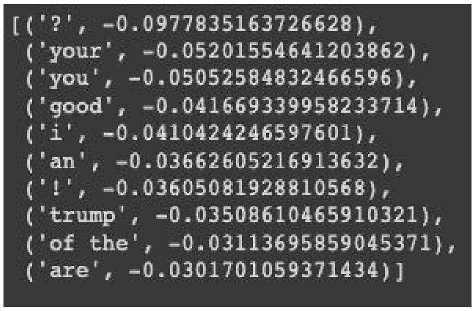
```
[('you can', -0.10671290439874159),
 ('?', -0.0977835163726628),
 ('not', -0.08642058458433238),
 ('it', -0.05929235116125474),
 ('in', -0.05630770540500851),
 ('good', -0.041669339958233714),
 ('i', -0.041042424659760 1),
 ('!', -0.03605081928810568),
 ('is a', -0.03520319516472407),
 ('trump', -0.03508610465910321)]
```

Black Toxic Weights False Positive



```
[('women', 0.3653249849250808),
 ('dead', 0.2918414414423233),
 ('is not', 0.23137142503814456),
 ('deserves', 0.18696855506861224),
 ('not', 0.1661786942450091),
 ('is a', 0.16608442158314227),
 ('human being', 0.14926827142119556),
 ('different', 0.14814848852047996),
 ('may', 0.1389241493294546),
 ('women .', 0.11622819750093244)]
```

Black Non-Toxic Weights False Positive

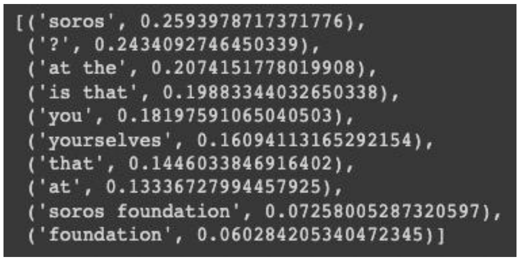


```
[('?', -0.0977835163726628),
 ('your', -0.05201554641203862),
 ('you', -0.05052584832466596),
 ('good', -0.041669339958233714),
 ('i', -0.041042424659760 1),
 ('an', -0.03662605216913632),
 ('!', -0.03605081928810568),
 ('trump', -0.03508610465910321),
 ('of the', -0.03113695859045371),
 ('are', -0.0301701059371434)]
```

Subgroup 2: Homosexual, Gay, or Lesbian

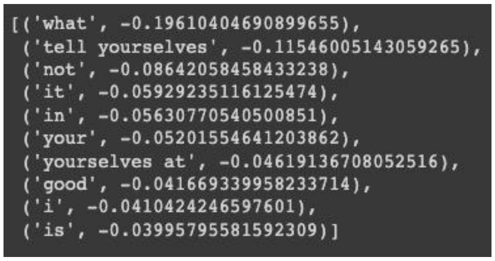
We did the same analysis as above with the homosexual subgroup. We showed the top results for positive weights and negative weights, for the true positive and false positive categories. The values for this subgroup are much lower than for the other subgroup, which also proves the point earlier that we made that because of the multiple words impacting this subgroup, the ADS might be getting confused. We also see more values that wouldn't make sense, such as trump being in the non-toxic weights top rankings of the homosexual comment subgroup. This further proves that the ADS is conflating identities as well as toxic/non-toxic comments.

H,G,L Toxic Weights True Positive



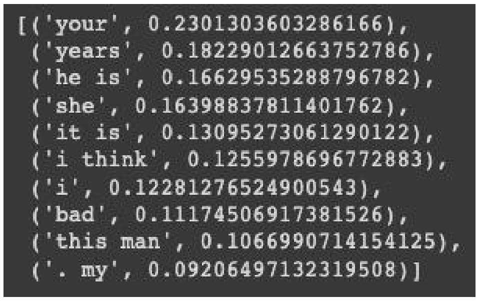
```
[('soros', 0.2593978717371776),
 ('?', 0.2434092746450339),
 ('at the', 0.2074151778019908),
 ('is that', 0.1988334403265O338),
 ('you', 0.1819759106504O503),
 ('yourselves', 0.16094113165292154),
 ('that', 0.1446033846916402),
 ('at', 0.13336727994457925),
 ('soros foundation', 0.07258005287320597),
 ('foundation', 0.060284205340472345)]
```

H,G,L Non-Toxic Weights True Positive



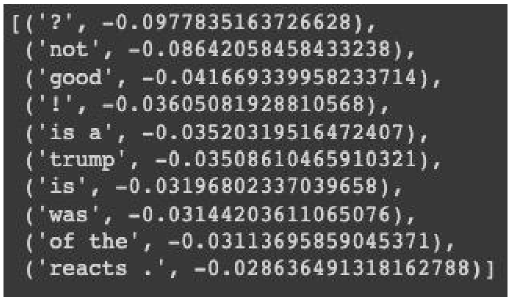
```
[('what', -0.19610404690899655),
 ('tell yourselves', -0.11546005143059265),
 ('not', -0.08642058458433238),
 ('it', -0.05929235116125474),
 ('in', -0.05630770540500851),
 ('your', -0.05201554641203862),
 ('yourselves at', -0.04619136708052516),
 ('good', -0.041669339958233714),
 ('i', -0.041042424659760 1),
 ('is', -0.03995795581592309)]
```

H,G,L Toxic Weights False Positive



```
[('your', 0.2301303603286166),
 ('years', 0.18229012663752786),
 ('he is', 0.16629535288796782),
 ('she', 0.16398837811401762),
 ('it is', 0.13095273061290122),
 ('i think', 0.1255978696772883),
 ('i', 0.12281276524900543),
 ('bad', 0.11174506917381526),
 ('this man', 0.1066990714154125),
 ('. my', 0.09206497132319508)]
```

H,G,L Non-Toxic Weights False Positive

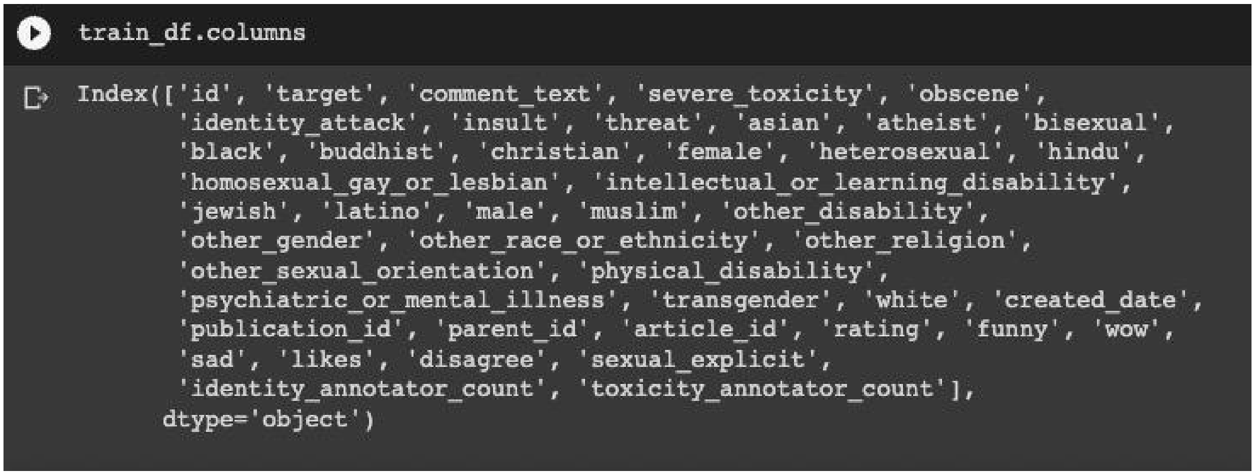


```
[('?', -0.0977835163726628),
 ('not', -0.08642058458433238),
 ('good', -0.041669339958233714),
 ('!', -0.03605081928810568),
 ('is a', -0.03520319516472407),
 ('trump', -0.03508610465910321),
 ('is', -0.03196802337039658),
 ('was', -0.03144203611065076),
 ('of the', -0.03113695859045371),
 ('reacts .', -0.028636491318162788)]
```

Summary

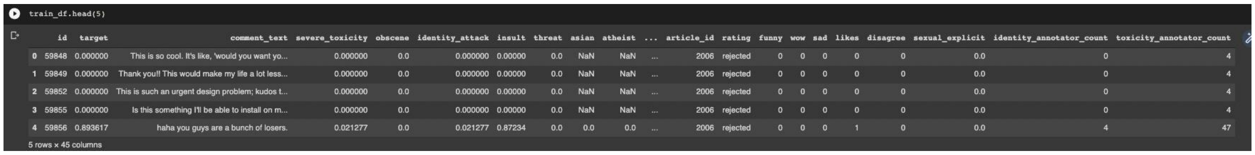
Data

I do believe that the data was appropriate for this ADS. It is organized in an extremely detailed manner, allowing those who read and analyze the data to clearly see that the target toxicity for the model is as well as each identity that is mentioned per comment. The only issue is, for logistic regressions, a binary y value is necessary for the model to work. In this case, the y value is a float value ranging from 0-1. This can be manipulated so that either there is a boolean value created determined by whether the target value is greater than or less than 0.5, and from there converted to a binary value of 0 or 1. The dataset might be easier to work with if this data was already processed to have a binary value for a target value.



```
train_df.columns
```

```
Index(['id', 'target', 'comment_text', 'severe_toxicity', 'obscene',
       'identity_attack', 'insult', 'threat', 'asian', 'atheist', 'bisexual',
       'black', 'buddhist', 'christian', 'female', 'heterosexual', 'hindu',
       'homosexual_gay_or_lesbian', 'intellectual_or_learning_disability',
       'jewish', 'latino', 'male', 'muslim', 'other_disability',
       'other_gender', 'other_race_or_ethnicity', 'other_religion',
       'other_sexual_orientation', 'physical_disability',
       'psychiatric_or_mental_illness', 'transgender', 'white', 'created_date',
       'publication_id', 'parent_id', 'article_id', 'rating', 'funny', 'wow',
       'sad', 'likes', 'disagree', 'sexual_explicit',
       'identity_annotator_count', 'toxicity_annotator_count'],
      dtype='object')
```



```
train_df.head(5)
```

	id	target	comment_text	severe_toxicity	obscene	identity_attack	insult	threat	asian	atheist	...	article_id	rating	funny	wow	sad	likes	disagree	sexual_explicit	identity_annotator_count	toxicity_annotator_count
0	59848	0.000000	This is so cool. It's like, 'would you want yo...	0.000000	0.0	0.000000	0.00000	0.0	NaN	NaN	...	2006	rejected	0	0	0	0	0	0.0	0	4
1	59849	0.000000	Thank you!! This would make my life a lot less...	0.000000	0.0	0.000000	0.00000	0.0	NaN	NaN	...	2006	rejected	0	0	0	0	0	0.0	0	4
2	59852	0.000000	This is such an urgent design problem; kudos t...	0.000000	0.0	0.000000	0.00000	0.0	NaN	NaN	...	2006	rejected	0	0	0	0	0	0.0	0	4
3	59855	0.000000	Is this something I'll be able to install on m...	0.000000	0.0	0.000000	0.00000	0.0	NaN	NaN	...	2006	rejected	0	0	0	0	0	0.0	0	4
4	59856	0.893617	haha you guys are a bunch of losers.	0.021277	0.0	0.021277	0.87234	0.0	0.0	0.0	...	2006	rejected	0	0	0	1	0	0.0	4	47

5 rows × 45 columns

Implementation

The accuracy of this model is approximately 94.3% and the false positive rate of the model is approximately 2.79%. As we mentioned earlier, the AUC of this model is also relatively low, which indicates that this model may not be extremely accurate with classification. For many of the different cases that we studied, the FN rates were quite large and the TP rates were quite small. This indicates the model has high specificity, but it has low sensitivity. This model is not effective enough to reach the original goal of being able to differentiate between comments that are toxic and comments that simply mention identity groups that are often attacked in comments. The model ends up classifying most comments as non toxic, and a few comments as toxic.

I also noted that the ADS has some mis-classifying issues when it comes to privileged vs unprivileged groups. There were a few situations in which the model would classify comments mentioning the identity “white” as toxic, even thought it is a privileged group. Comments that also didnt mention the words “homosexual”, “gay”, or “lesbian” are more likely to be classified as toxic than comments that do. This further demonstrates that this ADS has problems with classifying toxicity of comments with privileged/unprivileged groups fairly.

I used evaluation of false positive rates and false negative rates to analyze the incorrect classifications made by the model. We also used Mean Outcomes, Disparate Impact, and Equality of Opportunity to measure the fairness of the model’s classification. The fairness measures would be incredibly important if anyone wanted to use this in real world applications, because if it worked out to be biased against unprivileged groups, then it would cause incredible backlash for all stakeholders involved.

The model learned patterns of how to detect toxicity over time, such as detecting toxicity from words like “not” as well as strong punctuation that convey feelings. Through our analysis, it is also apparent that the model has learned to associate the mention of identity and nationality terms with toxicity.

Public Application

I would not be comfortable with deploying this ADS in the public sector or in the industry. The ADS has many issues, such as the fairness issues mentioned above, and the inability to accurately classify the toxicity of comments that mention multiple identity groups. Considering the high false negative rates of the model, the ADS would most likely identify many toxic comments as non toxic, and as a result, the ADS is not effective enough to roll out.

Improvements

There are many improvements that could be made to this model. The first improvement would be re reevaluating the minority scores for each comment, as they aren’t necessarily standard across the minority groups. Another improvement would be adding a metric to be able to determine the difference between toxic comments towards minority identities and non toxic comments towards minority identities. Another improvement would be to create a model that does not assume that the features are independent of each other, as many comments mention more than one identity.