AstraZeneca AI Challenge

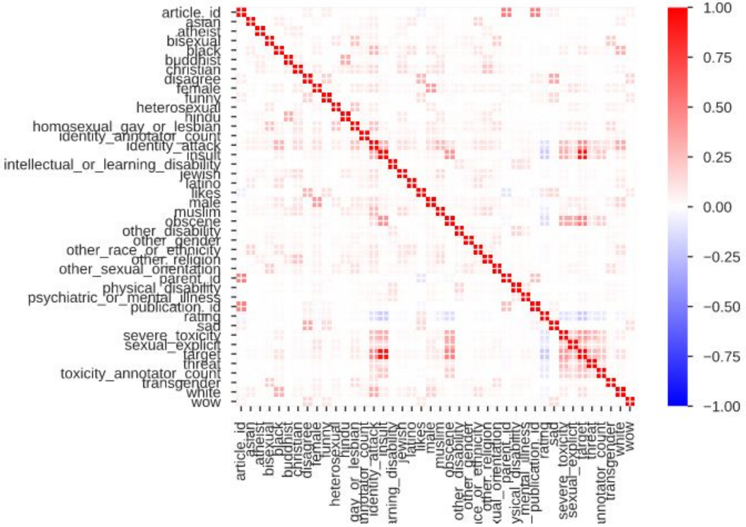
Team Name : CodingBeginner
Name : Rohith Srinivaas M
(SHA008208)
IIT Madras

JigSaw Unintended Bias in Toxicity Classification

- Objective : Rate the toxicity in a given data instance
- Target prediction is a fraction denoting the toxicity level in the comment thereby classifying it as a positive or negative comment.
- Subtype attributes for toxicity:
    - Severe_toxicity
    - Obscene
    - Threat
    - Insult
    - Identity_attack
    - sexual_explicit

Exploratory Data Analysis (EDA)

- Pandas profiler was used for quick EDA of the dataset.
- Important observations are given below:
  - **target** is highly correlated with **insult** with ρ = 0.928206624.
  - The overlap between **white** and **black** identity comments is high.
  - A large number of comments about the **Jewish** identity is toxic towards the **Muslim** identity



Approaches

- Using Logistic Regression
  - Using ELI5 to understand bias.
- Using textCNN
  - Using LIME to understand bias.

Metrics

- SubGroup AUC :
  - only to the examples that mention the specific identity subgroup
  - *A low value in this metric means the model does a poor job of distinguishing between toxic and non-toxic comments that mention the identity.*
- BPSN (Background Positive, Subgroup Negative) AUC :
  - To the non-toxic examples that mention the identity and the toxic examples that do not
  - *A low value in this metric means that the model confuses non-toxic examples that mention the identity with toxic examples that do not, likely meaning that the model predicts higher toxicity scores than it should for non-toxic examples mentioning the identity.*

Reference : Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification

Metrics

- BNSP (Background Negative, Subgroup Positive) AUC :
  - To the toxic examples that mention the identity and the non-toxic examples that do not.
  - *A low value here means that the model confuses toxic examples that mention the identity with non-toxic examples that do not, likely meaning that the model predicts lower toxicity scores than it should for toxic examples mentioning the identity.*

Reference : Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification

Results and Discussion

- Using Logistic Regression
  - Used TweetTokenizer from nltk package with Tfidvectorizer to get the word-vector.
  - Performed Logistic Regression with the obtained word-vector with class = 1 for target >= 0.5 and vice versa
  - Used ELI5 for model interpretation:

y=1 (probability **1.000**, score **10.195**) top features

Contribution?	Feature
+10.815	Highlighted in text (sum)
-0.621	<BIAS>

oh , bullshit your opinions are as worthless as a word from trump is a lie is that all you ve got , the worn out , pathetic and clearly wrong " sore loser " argument ? are the federal courts sore losers ? are all the ags , legal scholars and republican attorneys who have opined the eo is illegal sore losers ? the only sore losers i see are trump , his brown nosing supporters and his cabinet it a muslim ban and that is illegal

y=0 (probability **0.995**, score **-5.213**) top features

Contribution?	Feature
+4.641	Highlighted in text (sum)
+0.572	<BIAS>

exactly , so many liberal based media will deny it or manage to utilize this story as a means of distracting the ongoing trump investigation we will now see that the dnc ended up being the driving force in creating something that never was it getting real juicy now

Results and Discussion

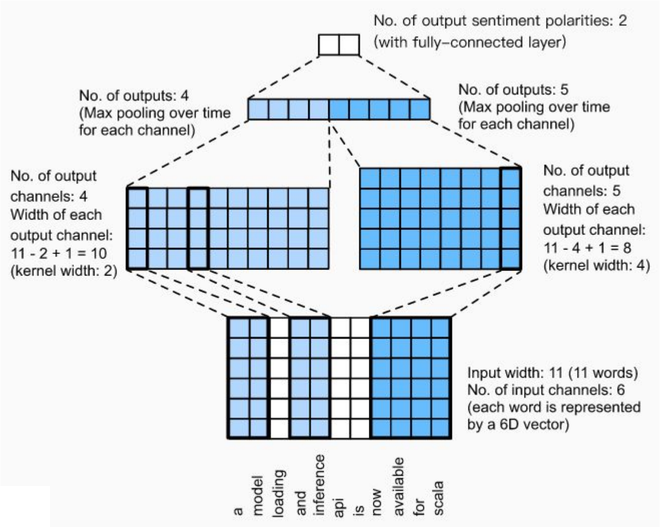
- Using TextCNN
  - One of the important hyperparameters in textCNN is the sequence length.

	Number of words	Sequence length	score
0	50000	150	0.9073
2	100000	150	0.9096
1	50000	300	0.9142
3	100000	300	0.9175

y=0 (probability **1.000**, score **-8.223**) top features

Contribution?	Feature
+7.550	Highlighted in text (sum)
+0.673	<BIAS>

they stood was that not enough ? now , can we please put this to bed ? stop the crying and move on < sheesh >



Improvements

- Using pretrained Word Embeddings like BERT, FastText, Glove to predict the toxicity and then use weighted sum of these vectors (with subgroup bias) to get the meta word-vector with minimum bias which can be used for further classification.

Thank You