Toxic Speech

Text Classification from Wikipedia comments

The Data

Description:

The dataset was created by Conversation AI an initiative by google.

The goal is to train a model to accurately predict different types of toxic speech

Scraped from Wikipedia.

Features:

159,571 comments from wikipedia ranging in length

Labels:

Toxic, Severe Toxic, Obscene, Threats, Insults, Identity hate

Grammar: does it even matter?

- Lemmatization
- Stemming
- Stop words
- Word tokenization
- Max features

...sort of

- Lemmatization
- Stemming
- Stop words
- Word tokenization
- Max features

Training Model

Processing the data:

- Tokenizing the data in a meaningful way
- Lemmatization and Stemming
- Reducing the number of features

Models:

Logistic Regression

Multinomial Naive Bayes

Understanding the Model

	non_toxic_coefs	toxic_coefs
case	-6.570649	NaN
two	-6.567903	NaN
policy	-6.545262	NaN
done	-6.522650	NaN
issue	-6.515988	NaN
...	...	...
bitch	NaN	-5.235449
suck	NaN	-5.101332
shit	NaN	-4.961371
fucking	NaN	-4.631957
fuck	NaN	-4.069232

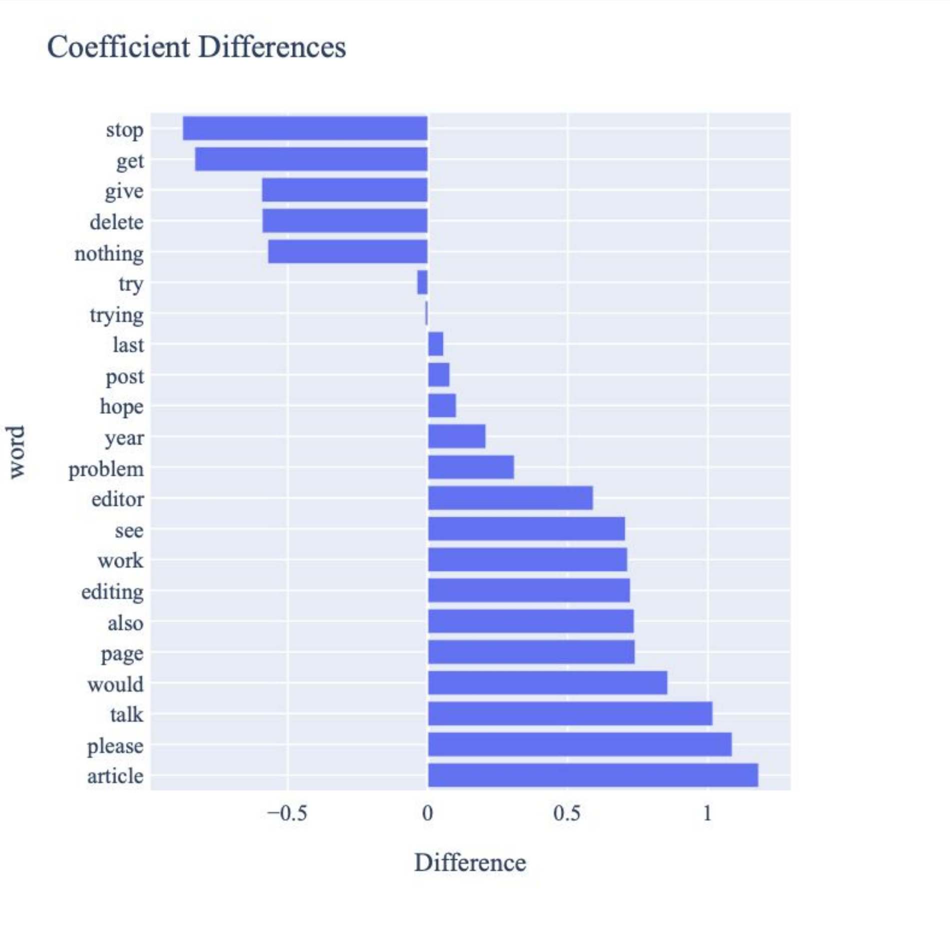
	non_toxic_coefs	toxic_coefs
case	-6.570649	-16.118096
two	-6.567903	-16.118096
policy	-6.545262	-16.118096
done	-6.522650	-16.118096
issue	-6.515988	-16.118096
...	...	...
bitch	-16.118096	-5.235449
suck	-16.118096	-5.101332
shit	-16.118096	-4.961371
fucking	-16.118096	-4.631957
fuck	-16.118096	-4.069232

Creating Stop Words

This process was important in identifying words that had a bearing on the model's prediction simply because they appear often in the dataset.

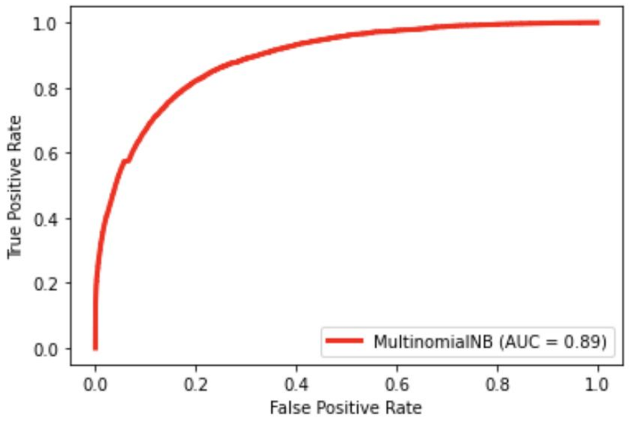
This was a helpful technique for identifying stop words.

It would be interesting to see how other models perform on the same words.



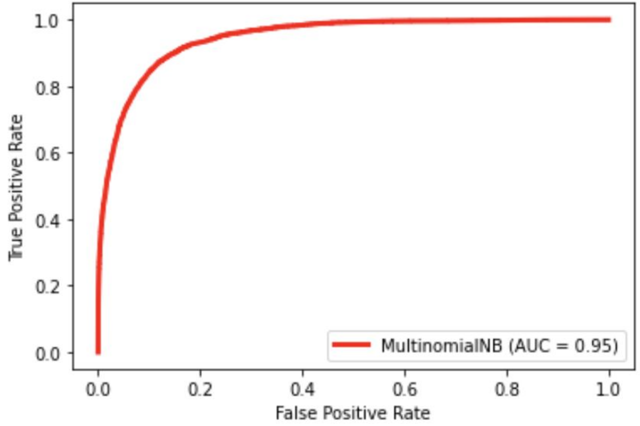
Before Stopwords

```
-------------------- MultinomialNB --------------------
TP:  1364
FP:  182
TN:  57706
FN:  4726

Precision: 0.8822768434670116
Recall: 0.22397372742200328
Accuracy: 0.9232861296070525
F1 Score: 0.357255107386066
Area Under ROC Curve: 0.8915241741938003
```



After Stopwords

```
-------------------- MultinomialNB --------------------
TP:  3589
FP:  1581
TN:  56307
FN:  2501

Precision: 0.6941972920696325
Recall: 0.5893267651888342
Accuracy: 0.9361968176560693
F1 Score: 0.6374777975133215
Area Under ROC Curve: 0.9468701678956976
```



Assessing Bias

Finding Bias:

Examine the coefficients from our trained model.

We can find the words with the highest coefficients for each class.

Understand the context of our data as well as the context in which those words appear and evaluate our model.

Examples of bias:

Words with high coefficients included, nazi, wiki, wikipedia, gay, vandalism, article, people, page, link etc.

Context is Everything

- Realizing my own bias when evaluating coefficients for the word nazi.
- Originally thought it was perfectly normal to see nazi in the toxic class
- Plenty of wikipedia comments about historical articles. Also who calls themselves a nazi before they commit hate speech?

Context 1:

> "pair of jew hating weiner *nazi* schmucks"

Context 2:

> "member of the american *nazi* party 91 further evidence of subcultures transversally as the *nazi's* did through a process"

Context 1:

> "fuck off u *gay* boy u r smelly fuck ur mum poopie."

Context 2:

> "the oppression of gay's and women the persecution of dissenting minorities"

Bias in our data

- Realizing my own bias when evaluating coefficients for the word nazi.
- Originally thought it was perfectly normal to see nazi in the toxic class
- Plenty of wikipedia comments about historical articles. Also who calls themselves a nazi before they commit hate speech?
- Realized my model was predicting the word gay as toxic

	word	percent labeled toxic
1	gay	0.548096
8	homosexual	0.383562
7	lesbian	0.303797
6	mexican	0.223776
11	transgender	0.206897
3	jew	0.162019
0	black	0.124219
10	feminist	0.118421
5	hispanic	0.115789
2	muslim	0.104839
4	jewish	0.091029
9	feminism	0.045455

Next steps

- Focus on ways to reduce bias
- Try n_grams to take context into consideration.
- Approach different parts of speech based tokenizations
- Understand the relationship between the words
- NMF, LSA

Thanks!