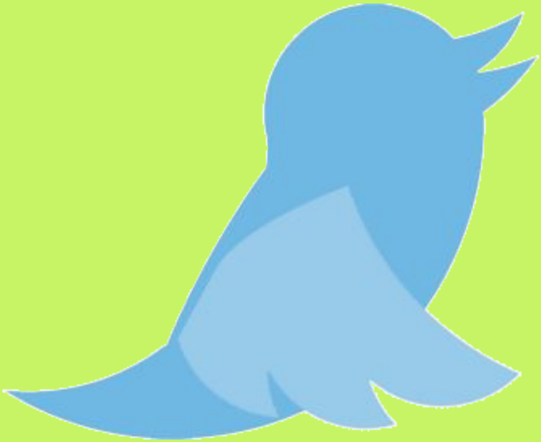


Hate Speech on Twitter

A Natural Language Processing Challenge

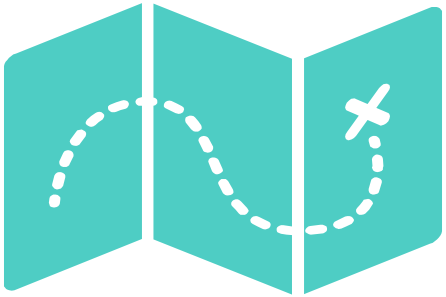
Ashley Steele

https://tinyurl.com/steeletwitternlp

Project Outline

Project Outline

1. About the Dataset
2. Exploratory Data Analysis (EDA)
3. Modeling our Data
4. Label Improvements and Future Implications



1.

About the Dataset

About the Dataset

Hate speech, as defined by the Oxford Constitutional Law website, is **"verbal or non-verbal communication that involves hostility directed towards particular social groups, most often on the grounds of race and ethnicity (racism, xenophobia, anti-Semitism, etc), gender (sexism, misogyny), sexual orientation (homophobia, transphobia), age (ageism), disability (ableism), etc."**

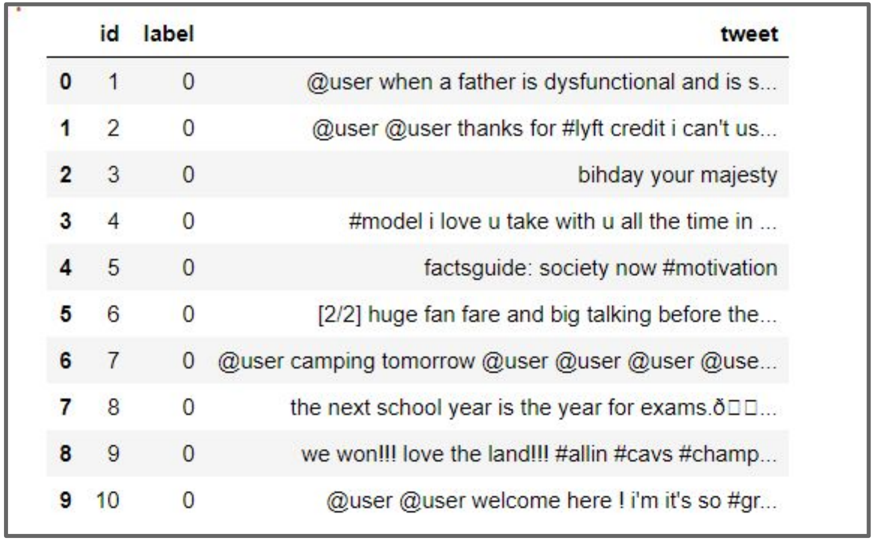
The aim of this dataset is to determine if a set of 30,000 tweets contains hate speech relating to sexism and racism in order to create predictive models to identify such language in the future.

This dataset is available online as a part of the Analytics Vidhya challenge series.

About the Dataset

Let’s break the data down:

- Over 30,000 tweets
- All tweets are in English
- Tweets are pre-labeled as either a 1 (containing hate speech) or a 0 (not containing hate speech)
- Each tweet has a unique id number
- Tweets themselves are noisy and often contain extra characters/symbols



	id	label	tweet
0	1	0	@user when a father is dysfunctional and is s...
1	2	0	@user @user thanks for #lyft credit i can't us...
2	3	0	bihday your majesty
3	4	0	#model i love u take with u all the time in ...
4	5	0	factsguide: society now #motivation
5	6	0	[2/2] huge fan fare and big talking before the...
6	7	0	@user camping tomorrow @user @user @user @use...
7	8	0	the next school year is the year for exams.ð□□...
8	9	0	we won!!! love the land!!! #allin #cavs #champ...
9	10	0	@user @user welcome here ! i'm it's so #gr...

Why is This Important?

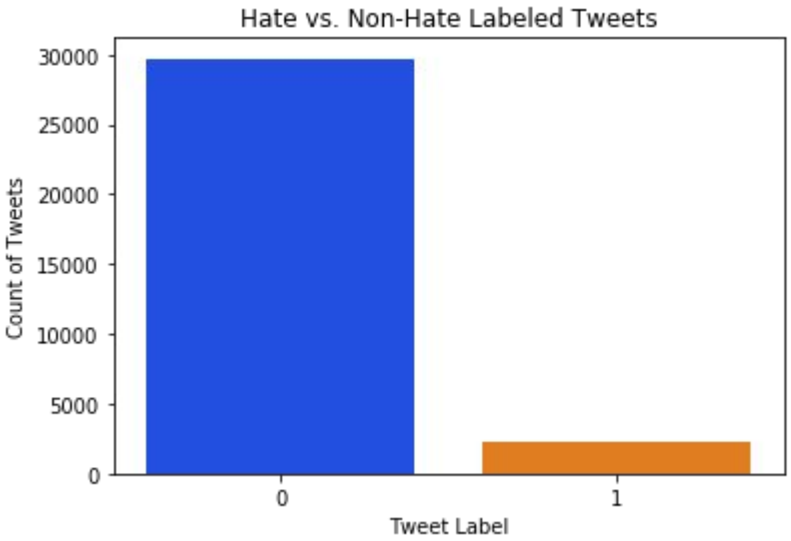
- Tweets containing hate speech can help predict hate crimes in the area they were generated in
- It is very difficult for social media platforms to police their user’s posts for hate speech
- Creating a predictive model for hate speech can automize detection
- Analyzing text for semantics (meaning) is a growing subset of data science called Natural Language Processing (NLP)

2.

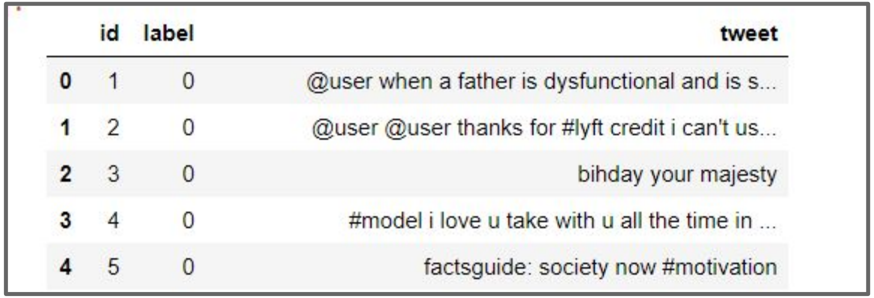
Exploratory Data Analysis (EDA)

Initial Data Exploration and Cleaning

What does the original data look like/tell us?



- Tweets are messy and have non-readable characters
- Less than 7% (2,242) of the total tweets (31,962) are labeled as hate speech



	id	label	tweet
0	1	0	@user when a father is dysfunctional and is s...
1	2	0	@user @user thanks for #lyft credit i can't us...
2	3	0	bihday your majesty
3	4	0	#model i love u take with u all the time in ...
4	5	0	factsguide: society now #motivation

Feature Engineering and More Exploration

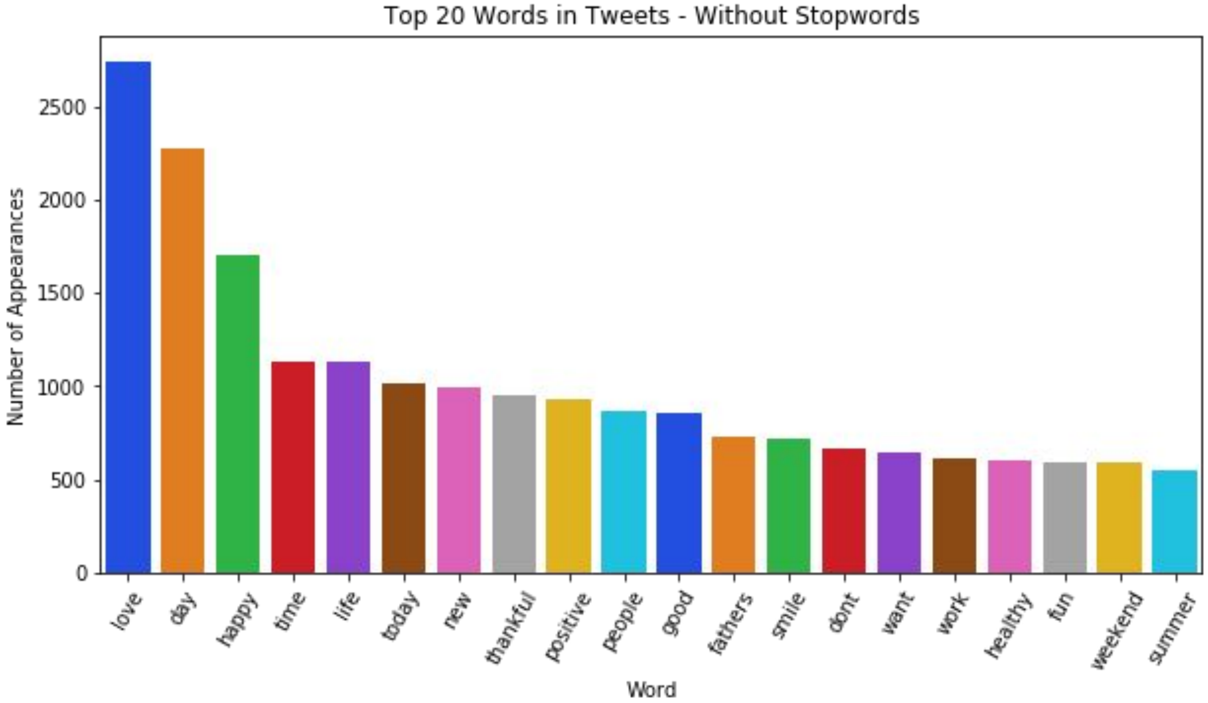
Does length of tweet relate to the tweet’s label?

Hate speech labeled tweets have an average length of 90.19 characters.

Non-hate speech labeled tweets have an average length of 84.33 characters.

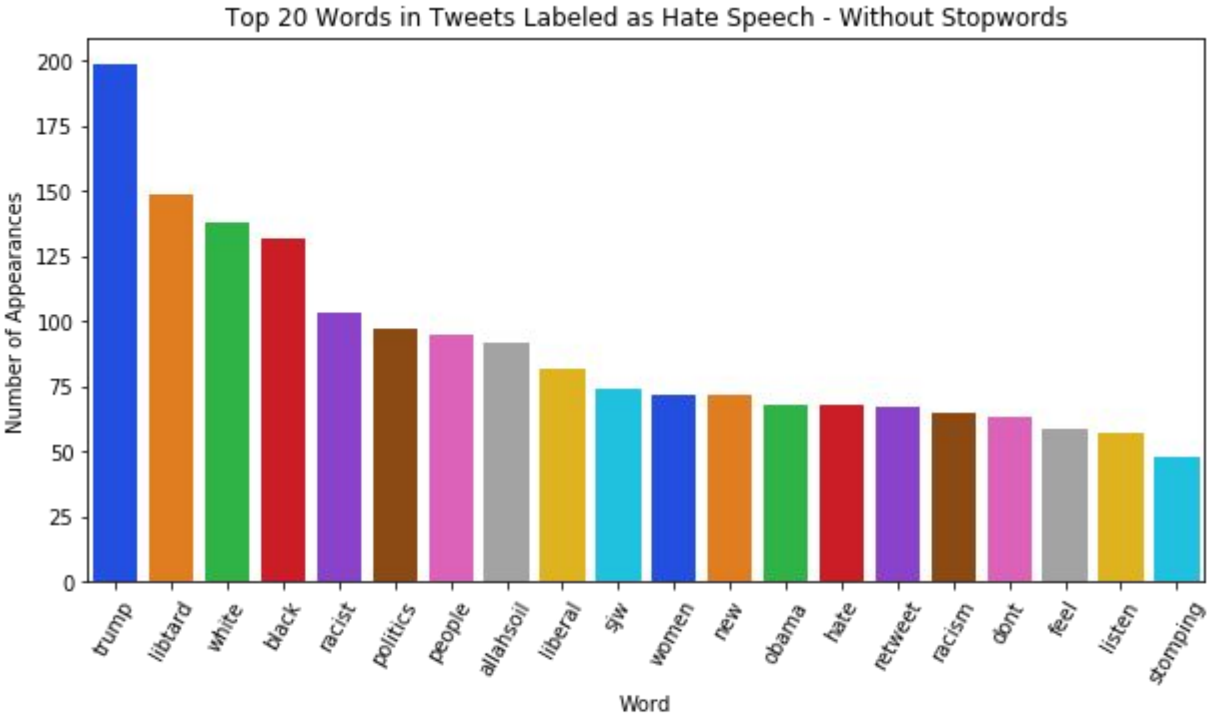
Cleaning Text and Keyword Identification

What are the most commonly occuring words in our data set?



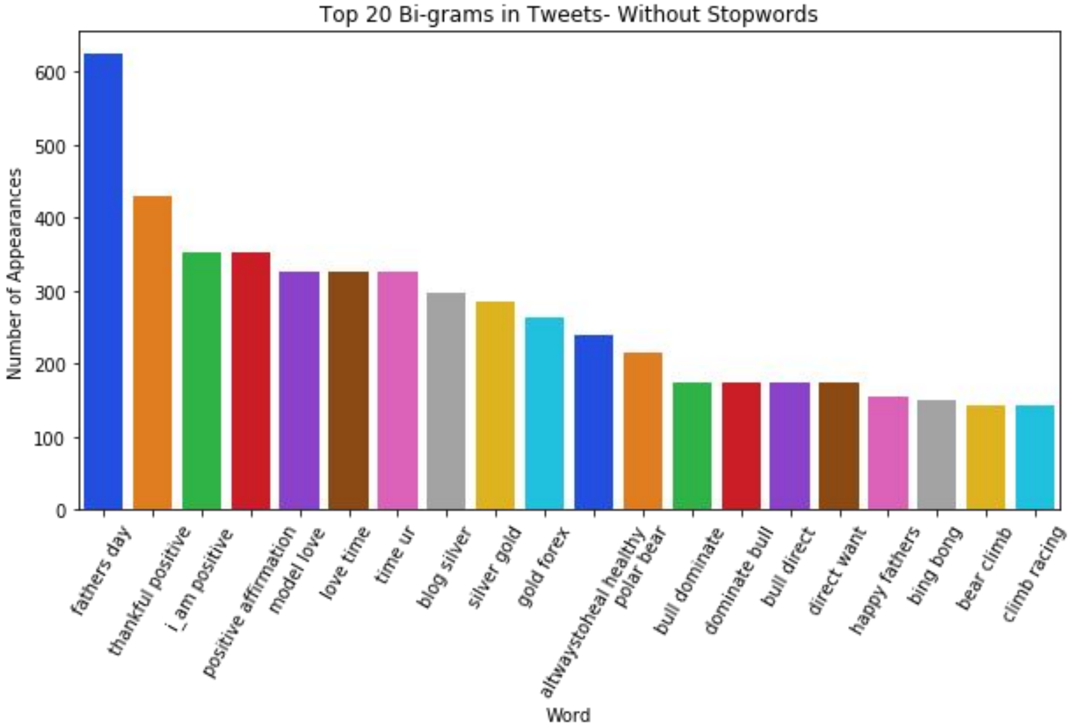
Cleaning Text and Keyword Identification

What are the most commonly occuring words in our data set?



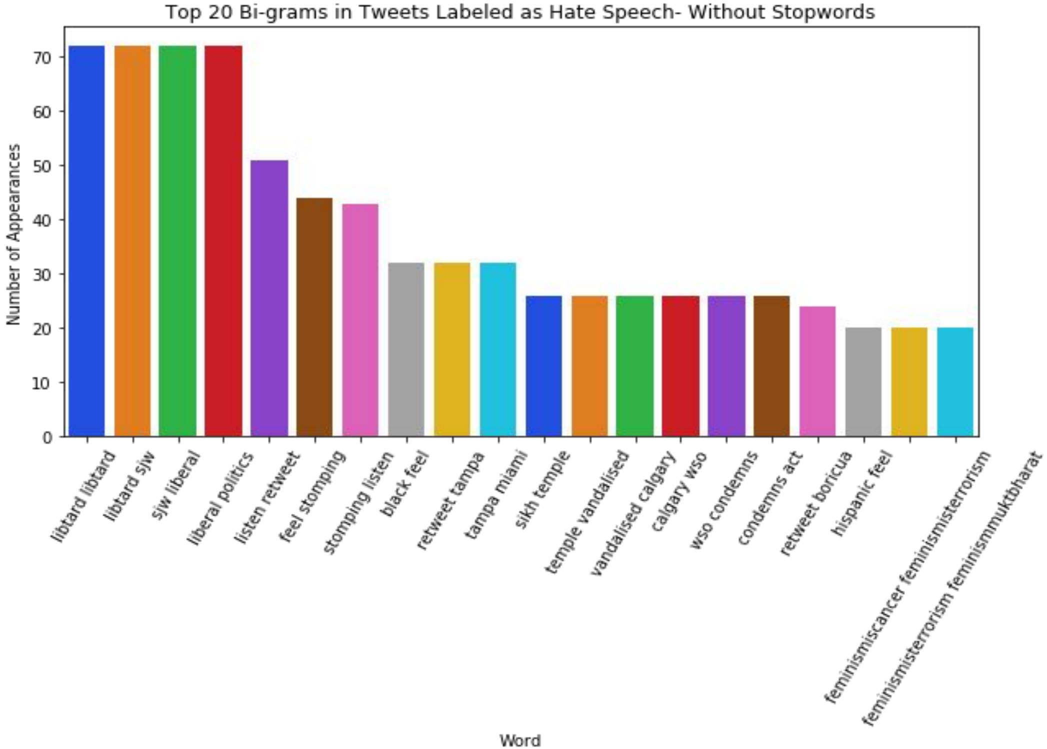
Cleaning Text and Keyword Identification

What are the most commonly occuring bigrams in our data set?



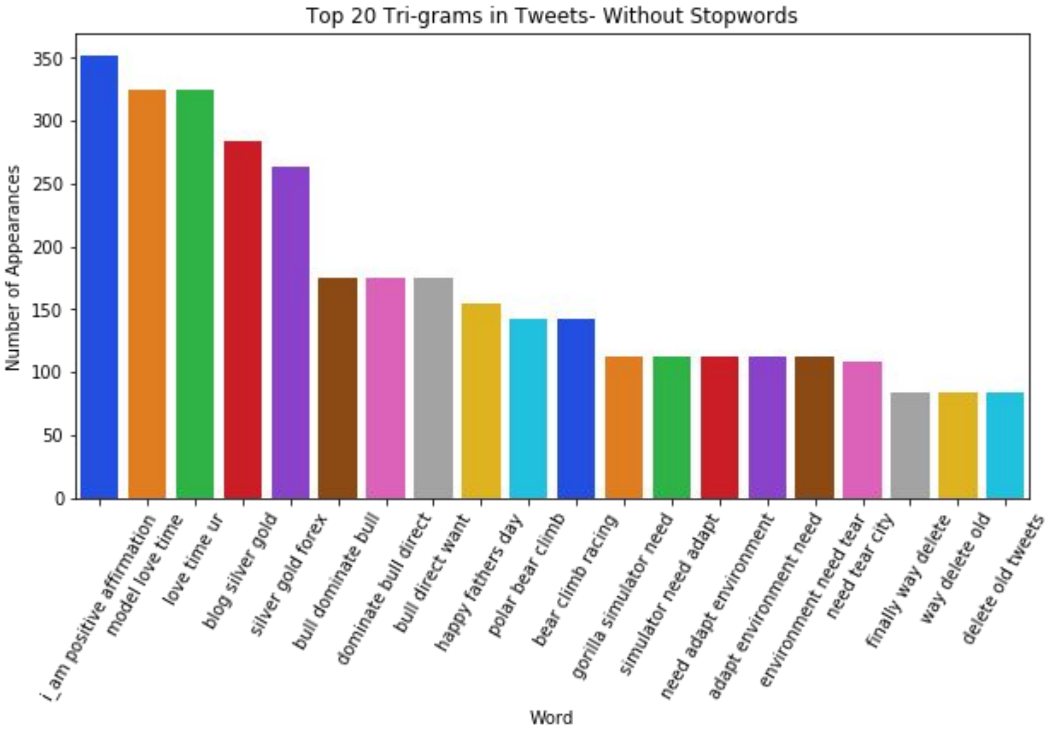
Cleaning Text and Keyword Identification

What are the most commonly occuring bigrams in hate-labeled tweets?



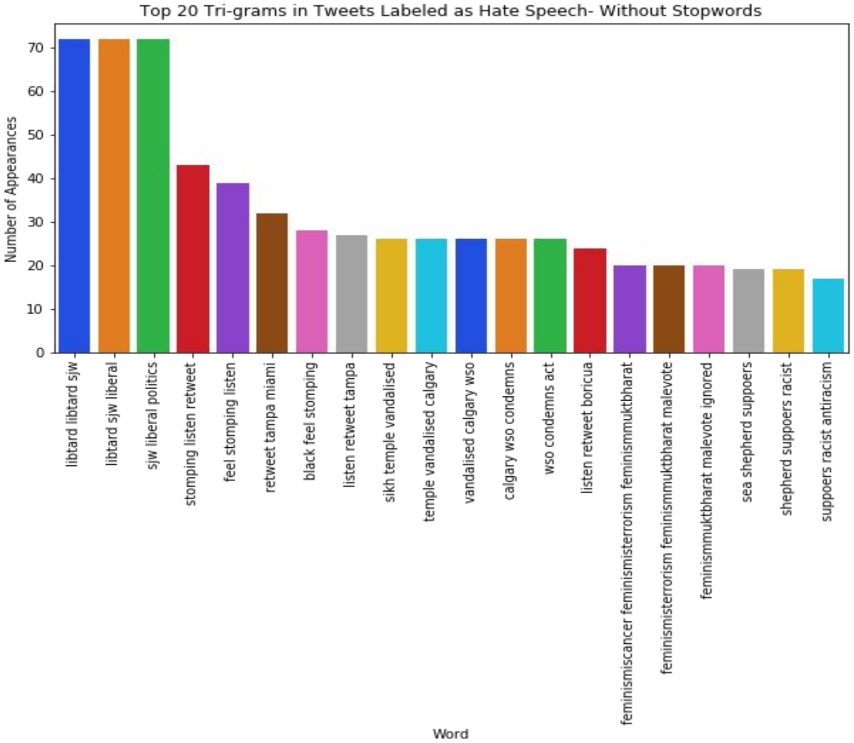
Cleaning Text and Keyword Identification

What are the most commonly occuring trigrams in our data set?



Cleaning Text and Keyword Identification

What are the most commonly occuring trigrams in hate-labeled tweets?



3.

Modeling Our Data

What is the Point of Our Model?

Research question: Can we use our prelabeled data set to predict if a tweet is hate speech or not?

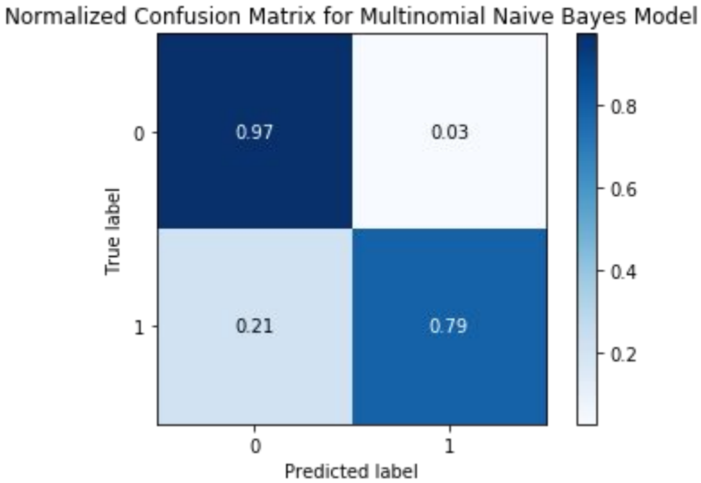
Model Set Up & Selection

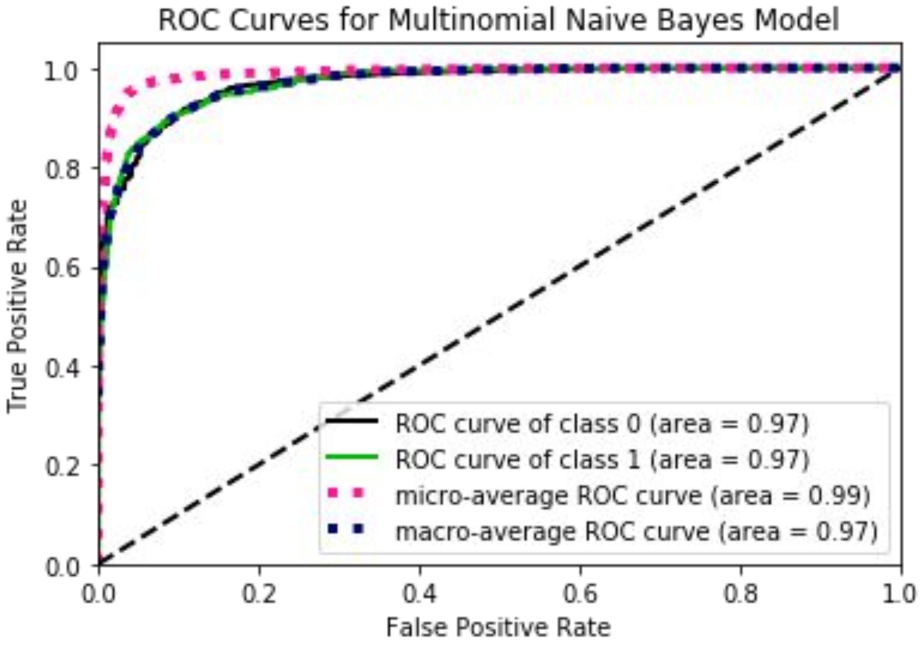
- ***X (feature) variables***: “bag of words” (the words from each tweet, seperated and counted)
- ***Y (target) variable:*** tweet label
- Our contestant are:
  - Multinomial Naive Bayes
  - Complement Naive Bayes
  - A Decision Tree
  - K-Nearest Neighbor (KNN)

Multinomial Naive Bayes

Why: Great with text & models with word count, fast, & simple

Results:


Normalized Confusion Matrix for Multinomial Naive Bayes Model

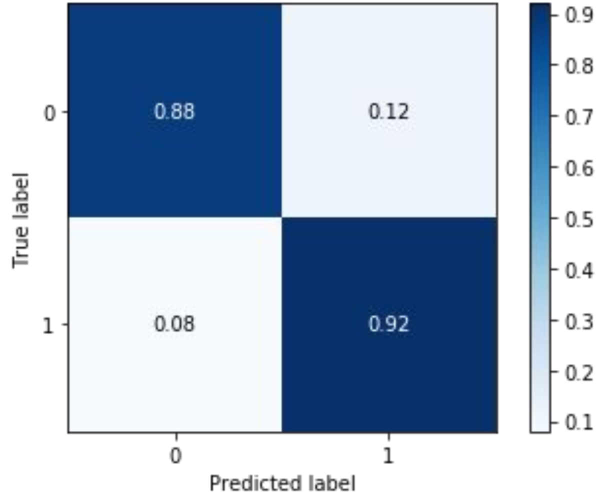

ROC Curves for Multinomial Naive Bayes Model

Complement Naive Bayes

Why: Great for imbalanced classes, simple, improvement on Multinomial NB

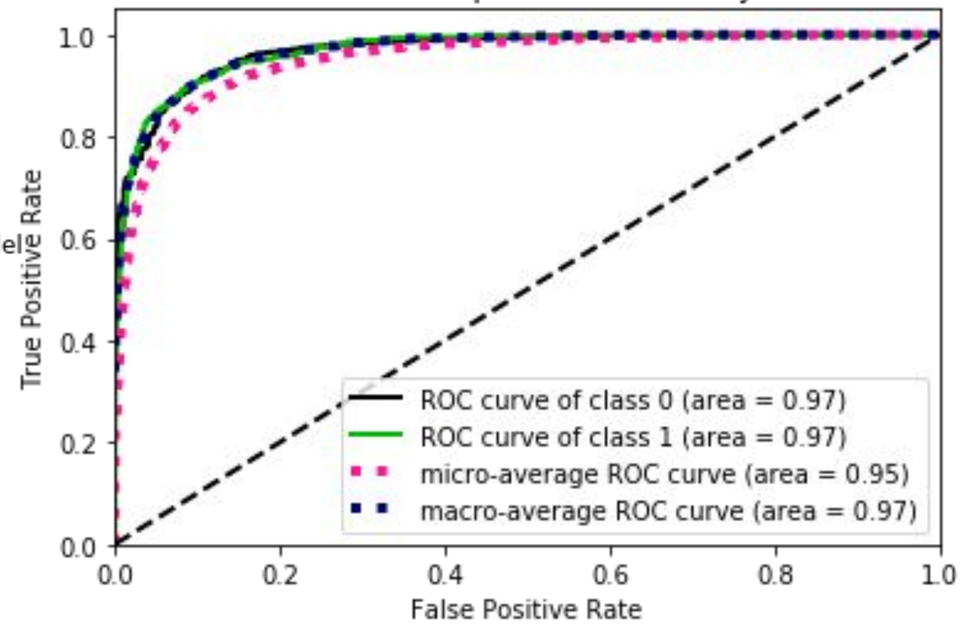
Results:

Normalized Confusion Matrix for Complement Naive Bayes Model



True label \ Predicted label	0	1
0	0.88	0.12
1	0.08	0.92

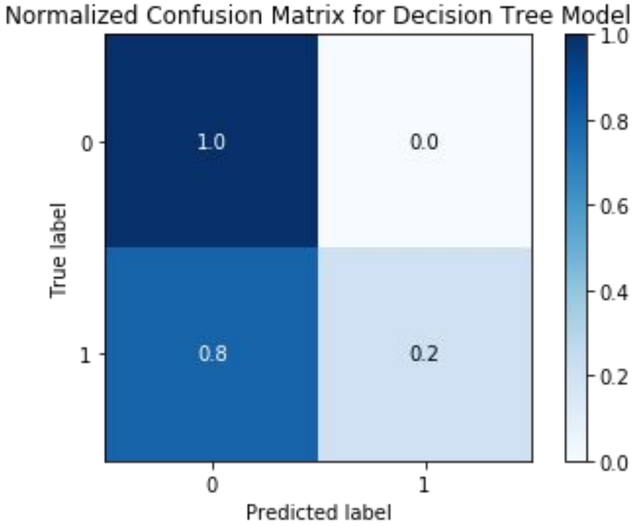
ROC Curves for Complement Naive Bayes Model

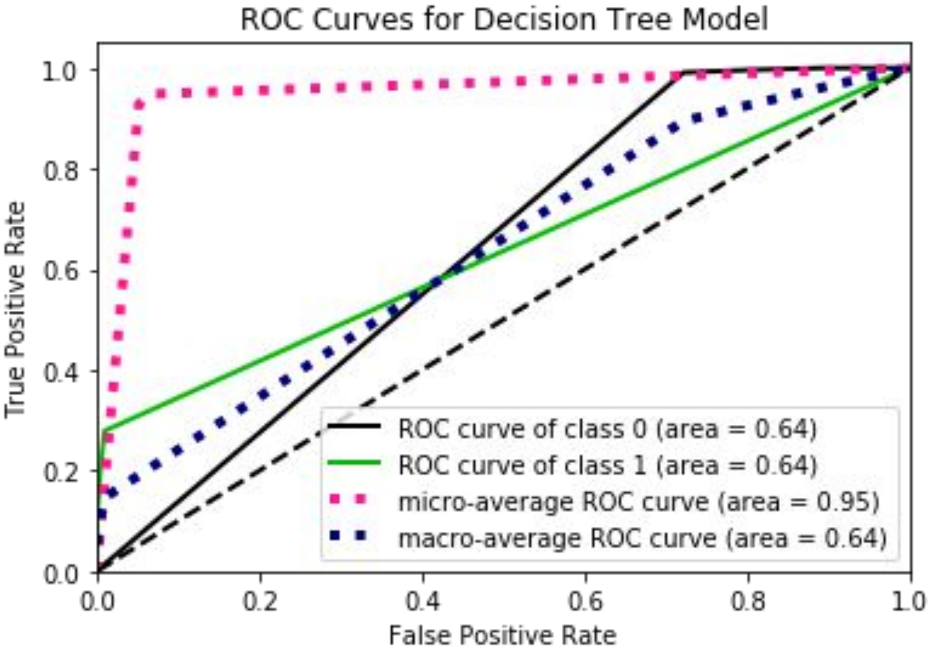


Decision Tree Model

Why: Simple to understand, ability to "see" what the model is doing

Results:

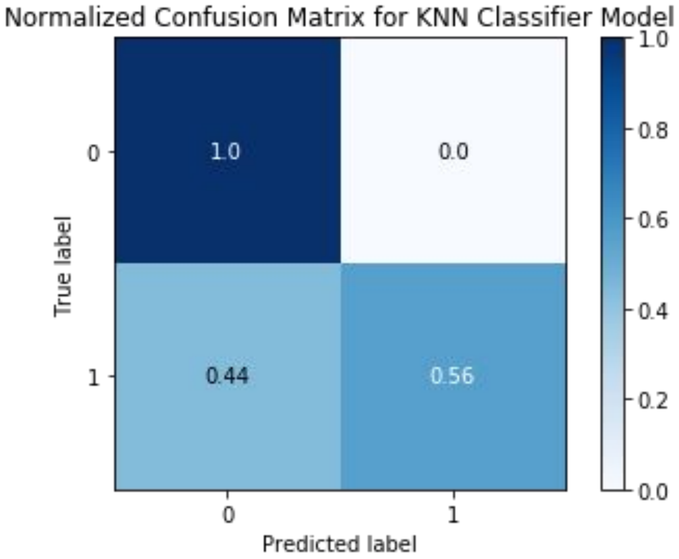


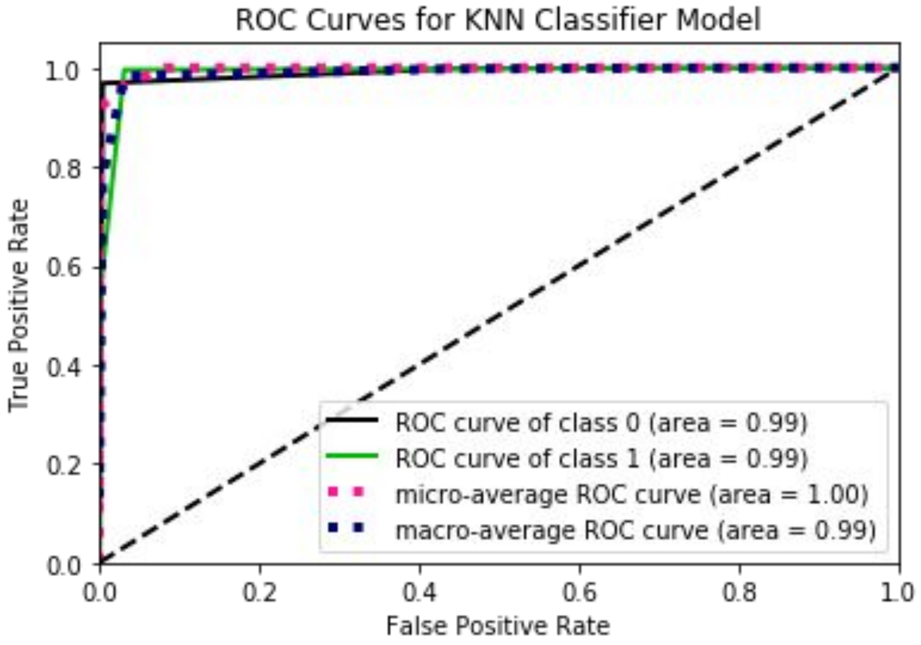


K-Nearest Neighbor Model

Why: More complex, works on "votes" of outcomes

Results:





Overall, What’s the Difference?

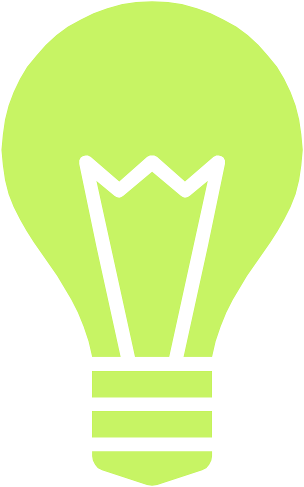
Model	Performance on 20% Holdout Group	Performance on Sample	Performance After Cross-Validation
Multinomial Naive Bayes	0.9460 or 95% explanation of variance in final outcome	0.9570 or 96% explanation of variance in final outcome	0.922 or 92% of explanation of variance in final outcome
Compliment Naive Bayes	0.8675 or 86%	0.8887 or 89%	0.8411 or 84%
Decision Tree	0.9399 or 94%	0.9397 or 94%	0.9342 or 93%
K-Nearest Neighbor	0.9419 or 94%	0.9633 or 96%	0.9381 or 94%

And the Winner is.....

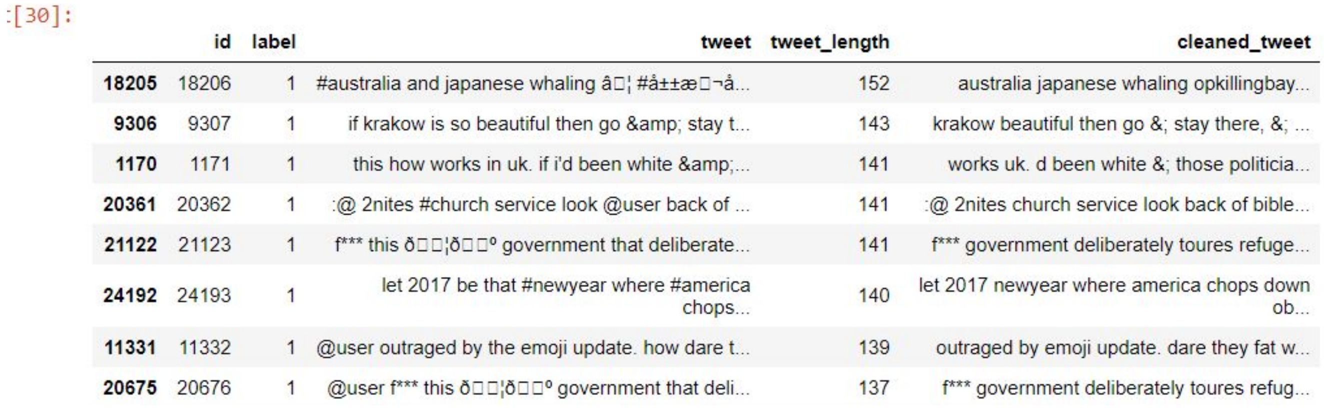
Multinomial Naive Bayes!

4.

Label Improvements

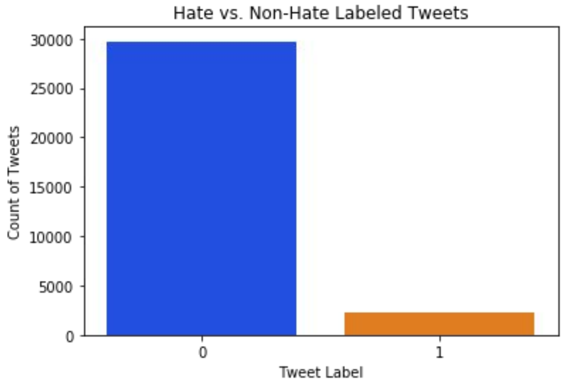


Did We Really Find Hate Speech?



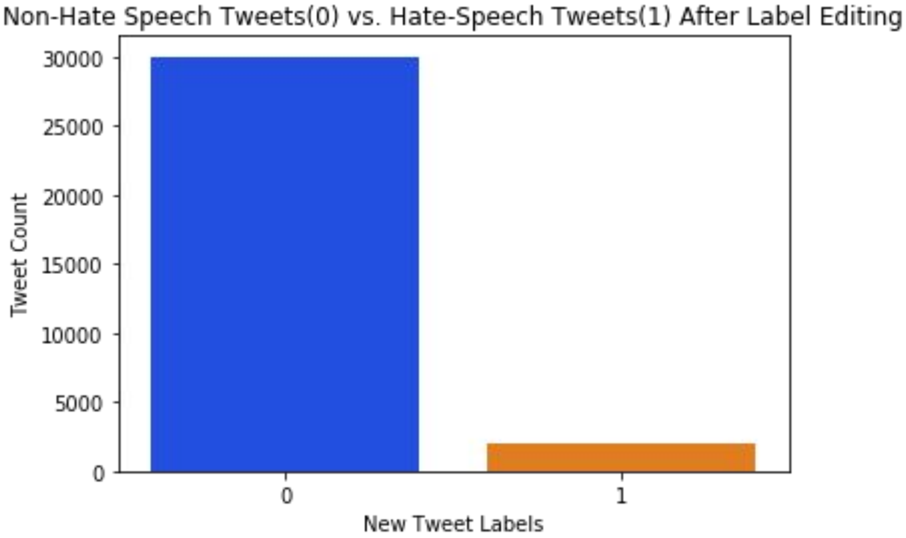
:[30]:

	id	label	tweet	tweet_length	cleaned_tweet
18205	18206	1	#australia and japanese whaling â□¦ #å±±æ□¬å...	152	australia japanese whaling opkillingbay...
9306	9307	1	if krakow is so beautiful then go &amp; stay t...	143	krakow beautiful then go &; stay there, &; ...
1170	1171	1	this how works in uk. if i'd been white &amp;...	141	works uk. d been white &; those politicia...
20361	20362	1	:@ 2nites #church service look @user back of ...	141	:@ 2nites church service look back of bible...
21122	21123	1	f*** this ð□□¦ð□□º government that deliberate...	141	f*** government deliberately toures refuge...
24192	24193	1	let 2017 be that #newyear where #america chops...	140	let 2017 newyear where america chops down ob...
11331	11332	1	@user outraged by the emoji update. how dare t...	139	outraged by emoji update. dare they fat w...
20675	20676	1	@user f*** this ð□□¦ð□□º government that deli...	137	f*** government deliberately toures refug...



Label Improvement and New Keyword Creation

- Created a new keyword list based on words identified in the Hurtlex
- Used keywords to create new labels
- Distribution of hate-labeled tweets decreased (from 7% to 6%)
- Keyword list is ever changing and improving



Future Implications

No really, why is this important to “future us”?

- Understanding of how computers can overfit NLP based on keywords (semantics and human touch important)
- “Experiments” such as this can help create a shared lexicon of hate speech to be used in NLP
- Improved quality of semantics detection in text can be transferred across industries and languages

Future Implications

Importance to “future me” (a.k.a. Main takeaways):

- Pay close attention to computing ability/time when modeling large amounts of data
- Experiment with different approaches, try whatever you can (without your computer exploding), and ask for help!
- Data science is so much more than just analytics!

Thanks!

Any questions?

You can find a copy of the original Jupyter notebook here!