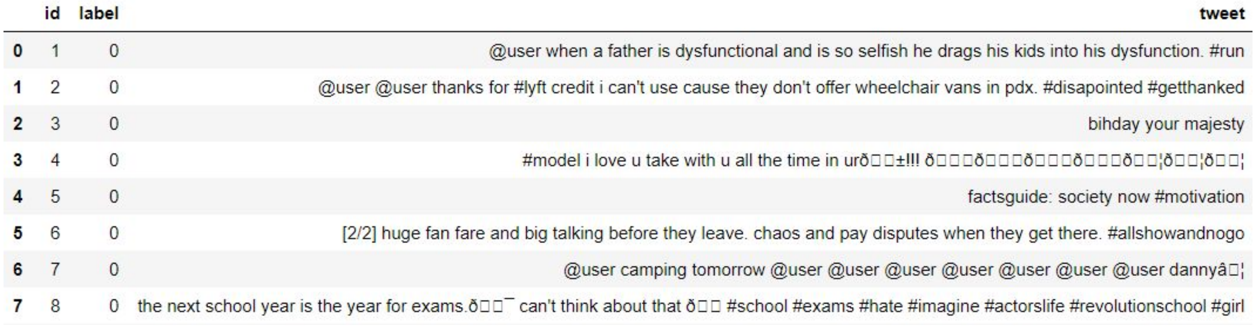
Classifying Hate Speech on Twitter: Milestone Report

Steve Donahue

Hate speech, trolling, and bullying on social media have become a serious problem in the last few years. From extra-national actors carrying out cyber campaigns exacerbate political differences to teenagers viciously abusing their peers, the internet has become a hostile place. The goal of this project is to develop a model which classifies racist and sexist language, using a previously labeled data set provided in a Vidhya Analytics challenge located here.

The data provided consists of unique index values, binary labels for hate speech vs normal speech, and the text of each tweet. All tweets have been represented as originally posted, except that the usernames have been removed to keep the posters anonymous. The raw data consists of about 32,000 training tweets and another 17,000 which serve as a validation set. The format for the raw data appears as follows.



	id	label	tweet
0	1	0	@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run
1	2	0	@user @user thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disapointed #getthanked
2	3	0	bihday your majesty
3	4	0	#model i love u take with u all the time in urð□□±!!! ð□□□ð□□□ð□□□ð□□□ð□□¦ð□□¦ð□□¦
4	5	0	factsguide: society now #motivation
5	6	0	[2/2] huge fan fare and big talking before they leave. chaos and pay disputes when they get there. #allshowandnogo
6	7	0	@user camping tomorrow @user @user @user @user @user @user @user dannyâ□¦
7	8	0	the next school year is the year for exams.ð□□¯ can't think about that ð□□ #school #exams #hate #imagine #actorslife #revolutionschool #girl

This project will clean, tokenize, and normalize the tweets using standard string processing methods and NLP procedures from the nltk module. The resulting processed tweets will be engineered into various features using the bag of words, ti-idf, word to vector, and document to vector methods. Then, SVM, Logistic Regression, Random Forest, and Light GBM classifiers will be trained on the data set and evaluated using the F1 metric. The top performing classifiers will then be further tuned for optimal classification.

The ultimate goal of this exercise is to produce a classifier which accurately determined which social media posts are racist or sexist. With a highly accurate model, it may be integrated into social platforms as an optional filter for a user's feed or a way to voluntarily prohibit inflammatory posts from being left on an individual user's wall.

It is the goal of this project to allow social media users the ability to customize their experience by electing not to entertain hate speech where they share their personal stories, images, and experiences.

Data Processing for EDA:

Several operations are applied to the raw tweets to turn them into usable, normalized, “tidy” tokens. First, we’ll define a function to edit the strings directly, based on boolean logic matching patterns we can specify.

```
def remove_pattern(input_txt, pattern):
    r = re.findall(pattern, input_txt)
    for i in r:
        input_txt = re.sub(i, '', input_txt)
    return input_txt
```

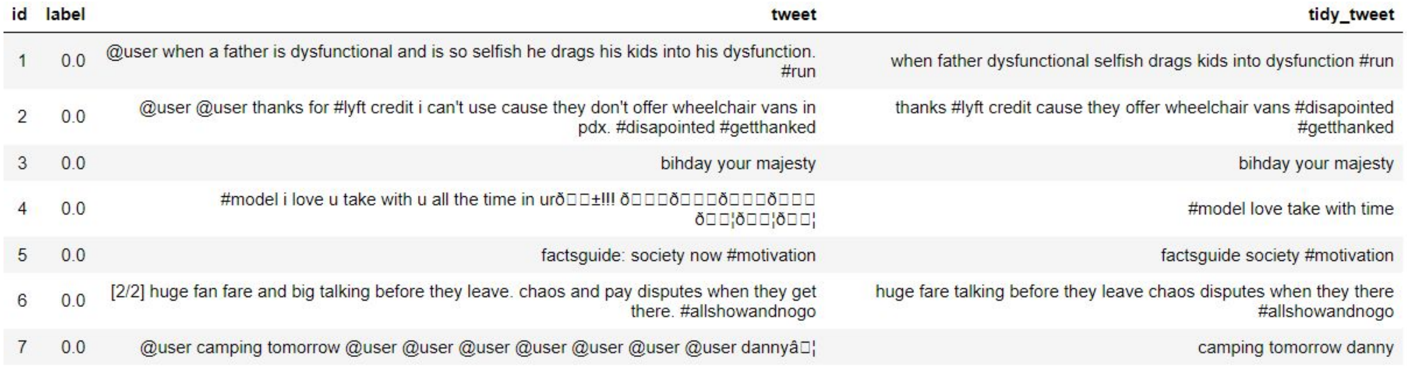
This function is then applied to a combined df of both test and train data to remove the @username strings, special characters, and all words of length 3 or less, using the following calls respectively.

```
combined['tidy_tweet'] = np.vectorize(remove_pattern)(combined['tweet'], "@[\w]*")
```

```
combined['tidy_tweet'] = combined['tidy_tweet'].str.replace("[^a-zA-Z#]", " ")
```

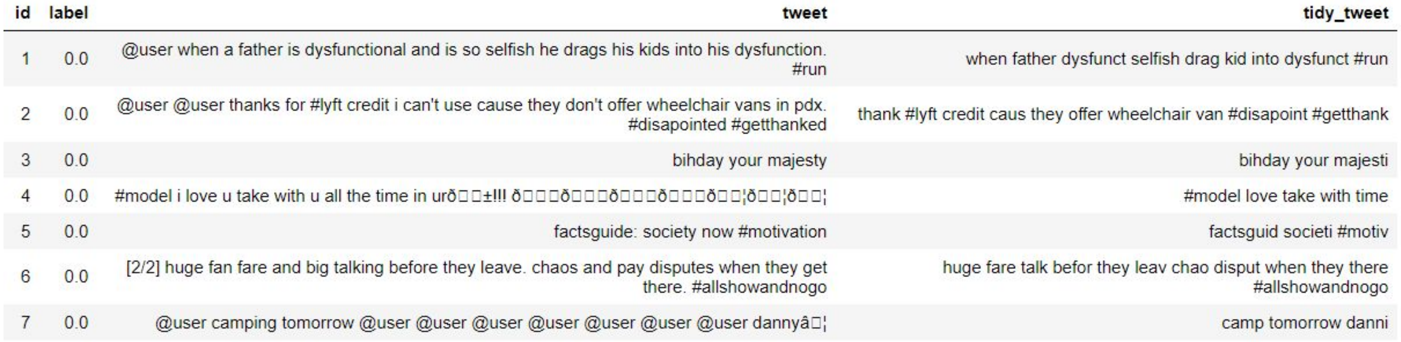
```
combined['tidy_tweet'] = combined['tidy_tweet'].apply(lambda x: ' '.join([w for w in  x.split() if len(w)>3]))
```

We obtain the following “tidy” dataset:



id	label	tweet	tidy_tweet
1	0.0	@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run	when father dysfunctional selfish drags kids into dysfunction #run
2	0.0	@user @user thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disapointed #getthanked	thanks #lyft credit cause they offer wheelchair vans #disapointed #getthanked
3	0.0	bihday your majesty	bihday your majesty
4	0.0	#model i love u take with u all the time in urð□□±!!! ð□□□ð□□□ð□□□ð□□□ ð□□¦ð□□¦ð□□¦	#model love take with time
5	0.0	factsguide: society now #motivation	factsguide society #motivation
6	0.0	[2/2] huge fan fare and big talking before they leave. chaos and pay disputes when they get there. #allshowandnogo	huge fare talking before they leave chaos disputes when they there #allshowandnogo
7	0.0	@user camping tomorrow @user @user @user @user @user @user @user dannyâ□¦	camping tomorrow danny

Next, we normalize and tokenize the words in each tweet. This means transforming various grammatical forms of a word to a single version. For instance, in tweet number 6 above, the word “talking” is transformed to “talk.” “This,” “those,” “these”, etc are transformed to “thi” which represents the normalized form for those words. These normalized values can now be counted and may contribute to feature generation for ML.



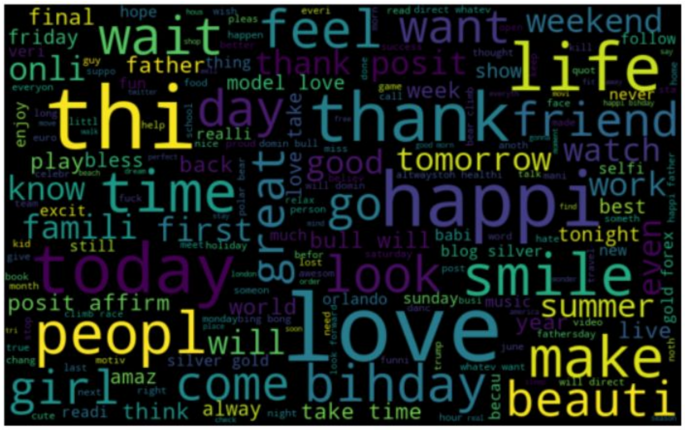
id	label	tweet	tidy_tweet
1	0.0	@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run	when father dysfunct selfish drag kid into dysfunct #run
2	0.0	@user @user thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disapointed #getthanked	thank #lyft credit caus they offer wheelchair van #disapoint #getthank
3	0.0	bihday your majesty	bihday your majesti
4	0.0	#model i love u take with u all the time in urð□□±!!! ð□□□ð□□□ð□□□ð□□□ð□□¦ð□□¦ð□□¦	#model love take with time
5	0.0	factsguide: society now #motivation	factsguid societi #motiv
6	0.0	[2/2] huge fan fare and big talking before they leave. chaos and pay disputes when they get there. #allshowandnogo	huge fare talk befor they leav chao disput when they there #allshowandnogo
7	0.0	@user camping tomorrow @user @user @user @user @user @user @user dannyâ□¦	camp tomorrow danni

The tidy data is then exported to a .csv file “processed_tweets” to be called by subsequent notebooks.

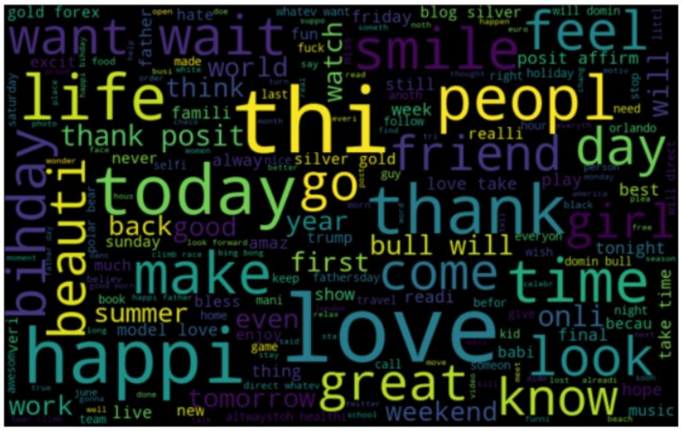
EDA and Statistics

A Wordcloud was chosen to illustrate the differences between normal tweets and those flagged as containing racist or sexist sentiments. A Wordcloud will show the word tokens most frequently occuring in the data fed to it.

The Wordcloud for the whole dataset:

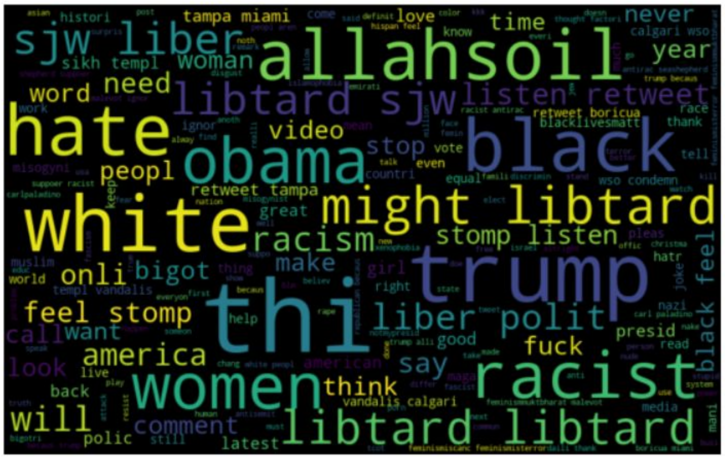


The Wordcloud for normal tweets:



There are few differences, due to the relatively small amount of hate speech overall.

The Wordcloud for hate speech takes on a much more political context.

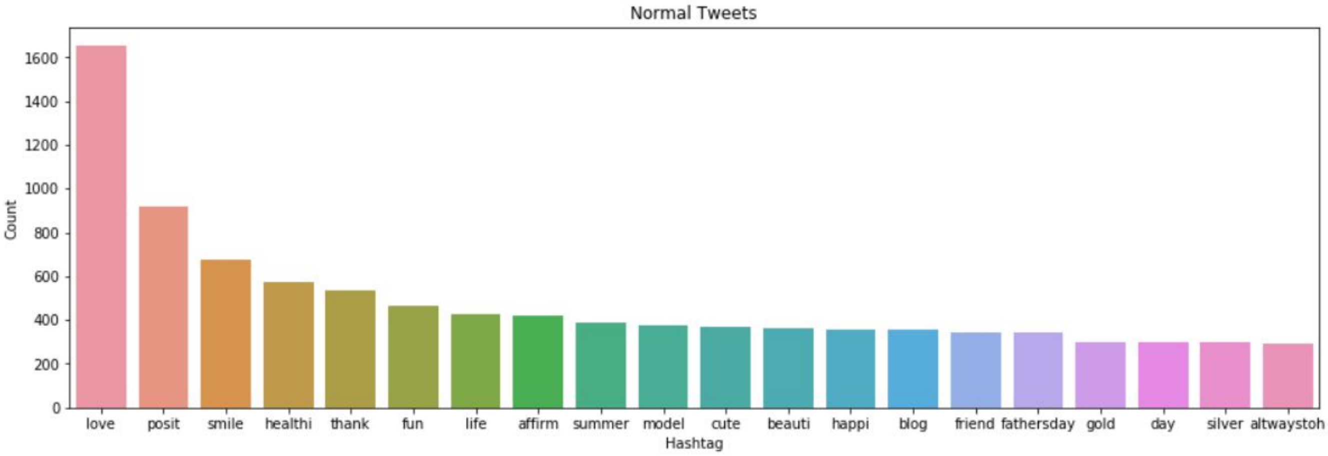


We can also develop some understanding of the difference between normal and abnormal tweets by examining the hashtag content across each. To do so, we define a function to extract the hashtags from each group and create a hashtag array.

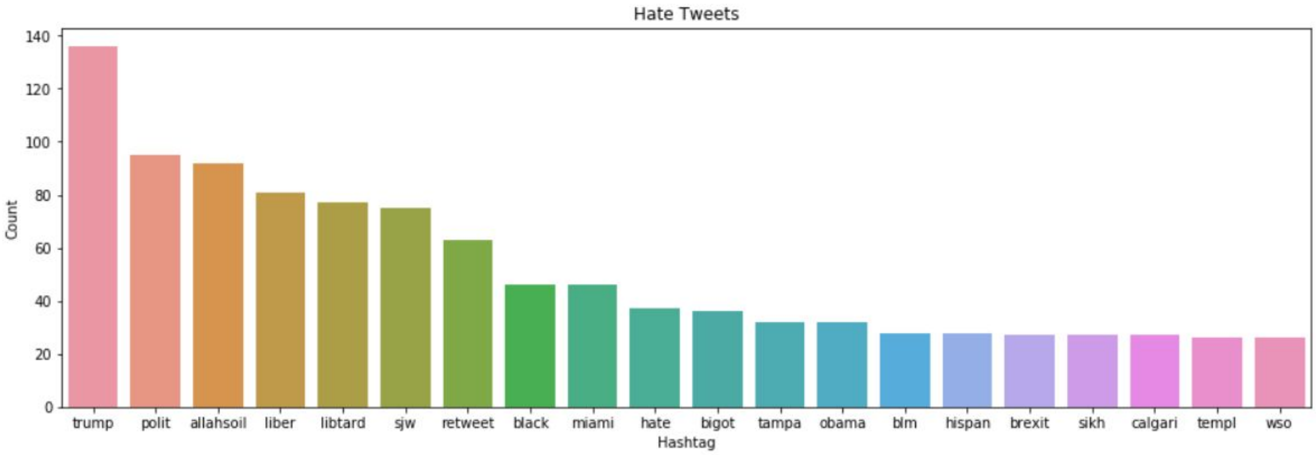
```
def hashtag_extract(x):
    hashtags = []

    # Loop over the words in the tweet
    for i in x:
        ht = re.findall(r"#(\w+)", i)
        hashtags.append(ht)
    return hashtags
```

Applied to each group and ranked, we can obtain bar graphs of the most frequently occurring hashtags in each sentiment group. Normal tweets are celebratory in nature.



Tweets in the hate speech group are predominantly political.



Conclusions:

There are significant differences between normal online speech and hate speech, measurable by the words appearing most frequently in each. Even a casual observer can see the difference clearly from the Wordclouds and predominant hashtags. While this difference might be a starting point to develop a filter, we cannot use keywords and hashtags alone. Censoring all tweets about Trump, Hillary, Miami, politics, and the color black would be too crude a tool and would stifle honest conversations. However, because a noticeable difference exists, it should be possible to develop classification tools via machine learning, which will be the next objective of this project.