Characterizing and Modeling Tweets within User Using Supervised and Unsupervised Machine Learning.

Justin Lin, Matthew Nemesure, Darshan Sundaram

Introduction

Twitter is a platform that allows users to share thoughts or comments in text format. There are few restrictions on what can be posted as long as it is within the 280 character limit. Once a user shares a "tweet" it is available for public viewing. If another user is friends with or "following" the OP (original poster), it will appear on their timeline – a running list of tweets that appear in the order of which they were posted. Given the nature of the timeline, it is very difficult to track the number of times a tweet is read (it can simply be scrolled past without reading). This presented a problem: the users needed a metric to evaluate the popularity of their tweet. The solution to this was the "favorite button" – a narcissists dream. While scrolling through his/her timeline, if a user stumbled upon a tweet that they particularly enjoyed, they could "favorite" the tweet. The count of favorites on each tweet provided the metric of interest for this study. The goal for our work was to utilize unsupervised machine learning to evaluate topical differences [within users] between tweets that procured more favorites vs. those that were less popular. Additionally, we used a supervised model to predict the performance of tweets given a variety of features.

Data

The data used for this project was gathered via the Twitter API. We pulled the 200 most recent tweets from Donald Trump (@realDonaldTrump), Alexandria Ocasio-Cortez (@AOC) and Elon Musk (@elonmusk). In addition to the content of the tweet itself, the API provided information about the number of favorites for each tweet, the time of day, and the number of followers for each account.

Methods

LDA – Unsupervised topic analysis

The first step was to get the data for each user using the Twitter API. Once the data was gathered, it was saved as a JSON file so the tweets used for analysis would not be continuously updated. The raw text of each tweet and favorite count was then extracted from the raw output of the API query. To clean the messy text, a variety of preprocessing was performed. All retweets were removed as they were not the original content of the poster. Additionally, all links were removed. This left only the original content typed and posted by the user. At this point, all stopwords and punctuation were removed to leave only content with a meaningful contribution to topic.

After preprocessing, each tweet was count vectorized into unigrams, bigrams, and trigrams that appeared in at least 2 posts. This was then run through LDA to get a probability distribution over 20 topics for each tweet.

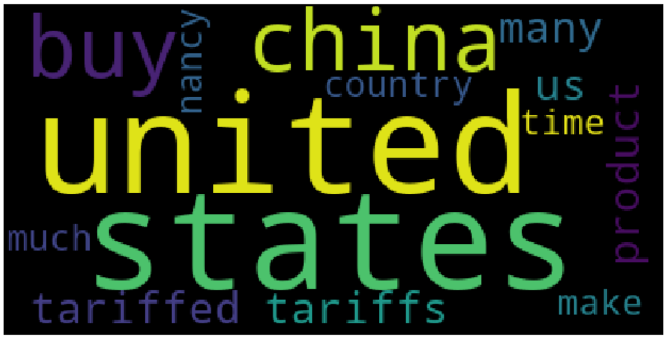
Once each tweet had its topic distribution, it was binned into low favorite count, medium favorite count or high favorite count. This binning was done on a percentile basis where the low bin was the bottom 33% and the top bin was the top 33%. The most common topic was then extracted for each group. This allowed for a qualitative analysis of topic differences between buckets.

Supervised Method for Predicting Buckets

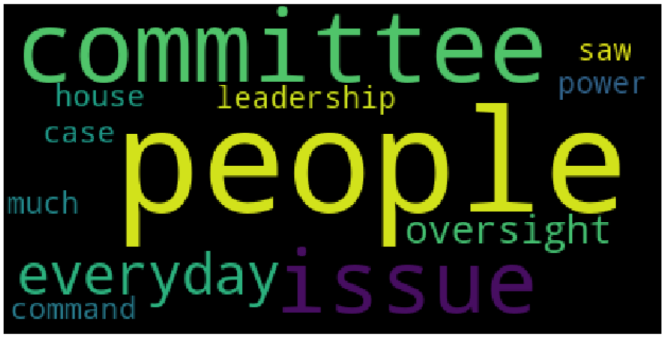
To perform supervised machine learning a data frame was built with features and outcomes and then split into training and validation. The outcome was the bin that a tweet belonged to making this a multiclass problem with three classes. Our features included the topic probability distribution for 20 topics from the LDA, the time of day for each tweet, the TF-IDF score and the day of the week. This data frame was built for each user and a variety of machine learning models including logistic regression, lightGBM, XGboost, and random forests were run for each data frame. The model was then tested on a withheld validation set to get predictive accuracy.

Results

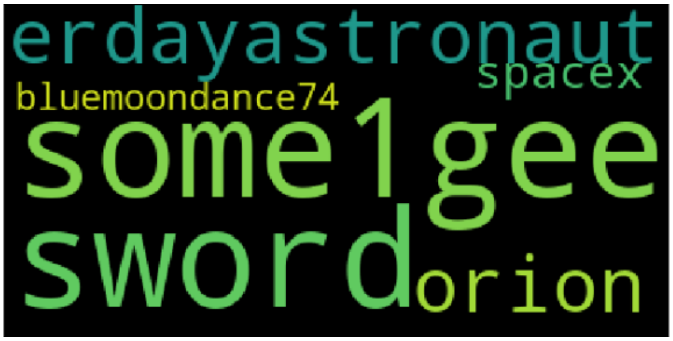
Topic for Trump's high favorite tweets



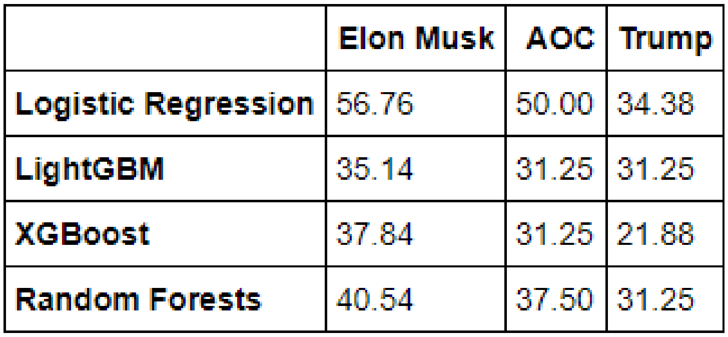
Topic for AOC high favorite Tweets



Topic for Elon's high favorite tweets



Supervised Machine Learning Accuracy



	Elon Musk	AOC	Trump
Logistic Regression	56.76	50.00	34.38
LightGBM	35.14	31.25	31.25
XGBoost	37.84	31.25	21.88
Random Forests	40.54	37.50	31.25

Donald Trump Full Results:

```
Most common topic of low favorite tweets:   total want total endorsement
complete fred republican complete total endorsement complete total foxnews
dems mueller congress crime Pennsylvania

Most common topic of medium favorite tweets:   china way people life wants
mueller deal mr like know report position bad last million

Most common topic of high favorite tweets:   china buy tariffs product us
many tariffed country make united nancy states much time united states
```

Alexandria Ocasio-Cortez Full Results:

```
Most common topic of low favorite tweets:   women leave living society tim
e nation us good policy think basic religious public power many

Most common topic of medium favorite tweets:   weeks law 12 use take would
year 12 weeks force forces days people pregnancy home etc

Most common topic of high favorite tweets:   people committee issue leader
ship everyday people oversight committee oversight case issues much house
power saw command
```

Elon Musk Full Results:

```
Most common topic of low favorite tweets:   tesla amp hours nozzle tesla y
es alex oha alex oha vacuum yes years close though would much

Most common topic of medium favorite tweets:   yes nichegamer spacex world
andscience spacex yes fischer fischer worldandscience course insideevs pun
ishedfranc83 spacexstarlink spacex punishedfranc83 spacexstarlink punished
franc83 spacexstarlink spacex spacexstarlink time

Most common topic of high favorite tweets:   erdayastronaut spacex erdayas
tronaut spacex some1gee erdayastronaut some1gee erdayastronaut spacex swor
d some1gee erdayastronaut some1gee orion sword sword orion sword some1gee
sword some1gee orion bluemoondance74 bluemoondance74 orion sword bluemoond
ance74 orion
```