Sentiment Analysis of Twitter Data for Prediction of Presidential Candidates

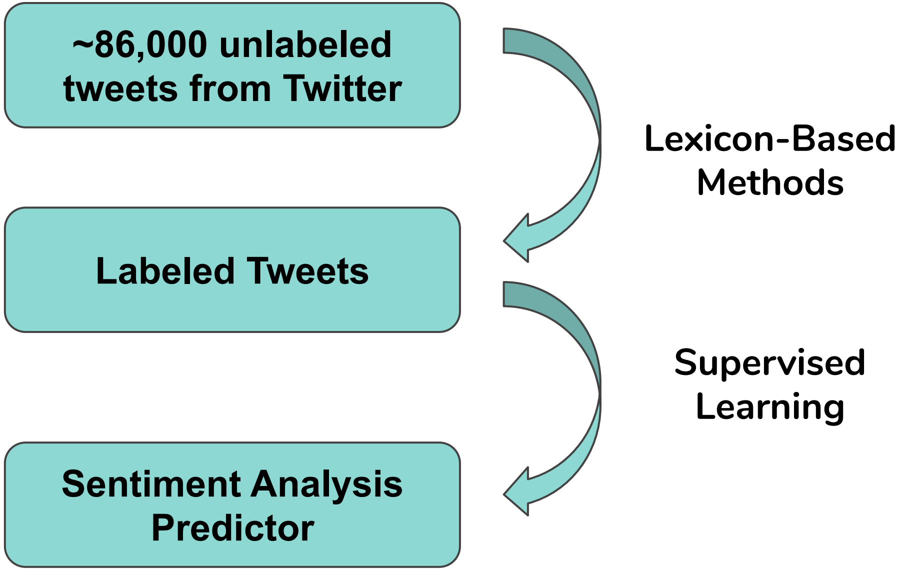
Dr. Deepti Srivastava Tilly

Many of the polls were wrong in predicting the outcome of the presidential election. Can we build a better predictor?

Who is running for president?

- Due to the number of Democratic candidates running for election this year, the scope of this project will be limited to the top 3 polling candidates.
- Using Twitter's API, tweets referencing the following candidates were extracted:
  - Bernie Sanders
  - Elizabeth Warren
  - Joe Biden

Proposed Workflow



~86,000 unlabeled tweets from Twitter

Lexicon-Based Methods

Labeled Tweets

Supervised Learning

Sentiment Analysis Predictor

Data cleaning involved dropping duplicate tweets and unnecessary columns and cleaning tweets to only remove noise.

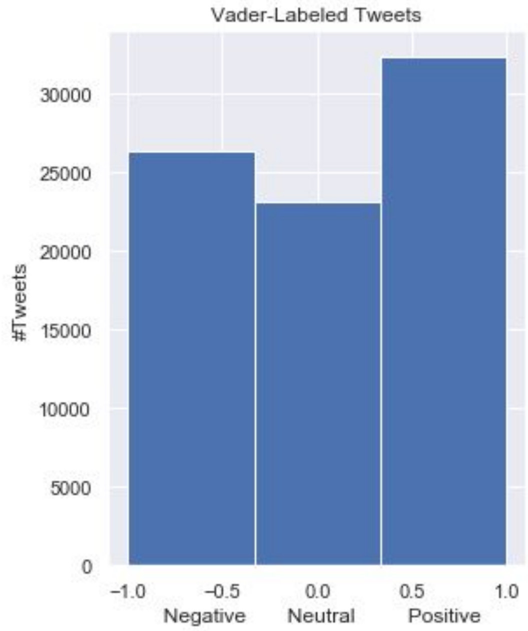
Dropping Duplicate Tweets and Unnecessary Data

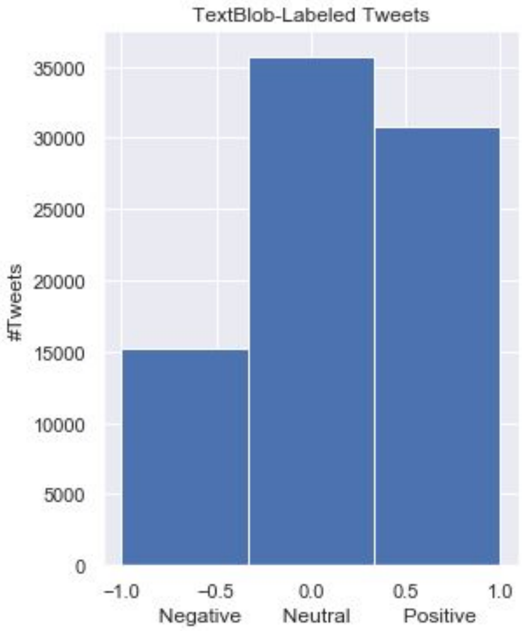
- ~5,000 duplicate tweets were found and removed, resulting in 81,633 tweets.
- Twitter's API returns a lot of extraneous data - only the tweet ID and text columns were retained.
- Each tweet was tagged as "sanders", "warren", or "biden" to easily identify which candidate's name was used to obtain the tweet.

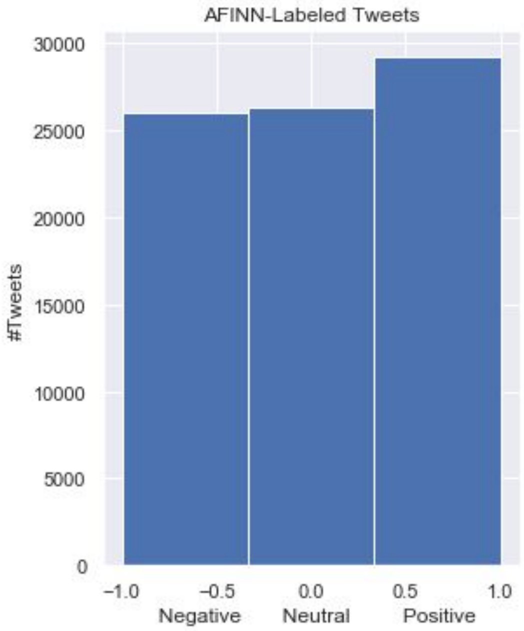
Cleaning Tweets

- Removing HTML tags using BeautifulSoup (e.g. &amp, &quot etc.)
- Removing @mentions
- Removing URLs
- Expanding Contractions ("don't" to "do not", "I'm" to "I am" etc.)
- Removing special characters (punctuation, numbers, #s)
- Removing extra white space and new line breaks
- Lemmatization
- Removal of stop words

In order to use a supervised learning classifier, the data was labeled using lexicon-based (i.e. dictionary) methods.

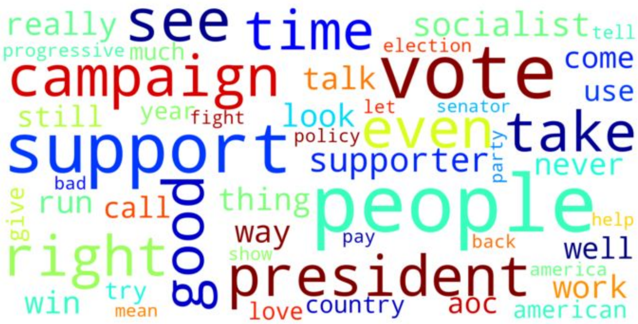






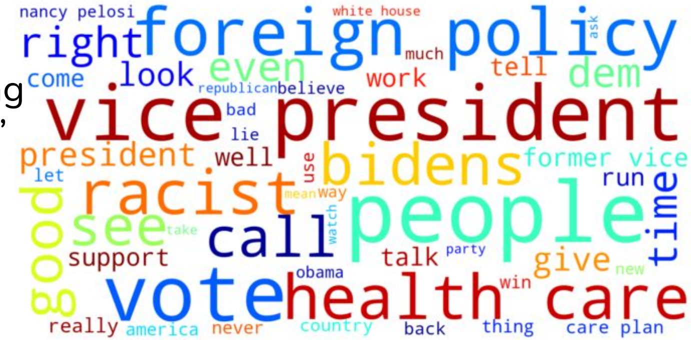
By comparing the performance with ~2000 manually labeled tweets, using a majority label was found to yield the highest accuracy.

EDA in the form of WordClouds revealed some interesting topics being discussed about the candidates.



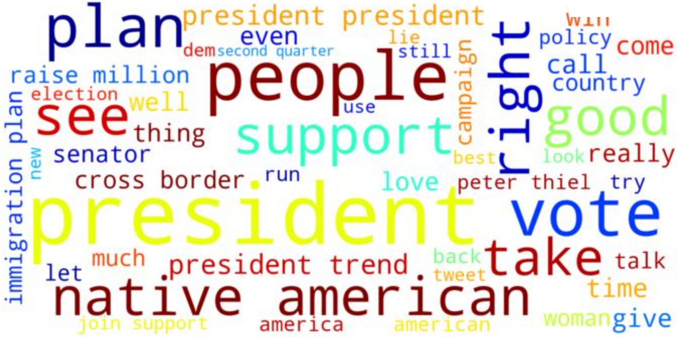
Warren:

- Warren's gaffe about calling herself a "native American" is being discussed.



Sanders:

- People are talking about his campaign and pledging support.



Biden:

- People are talking about Biden's healthcare and foreign policy. He is also being referred to as a racist.

Both regular BOW and TF-IDF models were used for feature extraction and 3 supervised learning models were tested. The Logistic Regression model was found to be the best performer.

	Naive Bayes Accuracy	Random Forest Accuracy	Logistic Regression Accuracy
BOW	0.75	0.81	0.86
TF-IDF	0.75	0.78	0.85

	Naive Bayes F1 Score	Random Forest F1 Score	Logistic Regression F1 Score
BOW	0.75	0.81	0.86
TF-IDF	0.74	0.8	0.85