Intro:

The only certainty in life is that nothing is certain. The presidential race in 2016 embodied this statement; most polls showed Hillary Clinton winning in a landslide with some indicating a 90% probability of victory. Unfortunately, polling is notoriously difficult and trying to procure a sample encompassing the U.S. population is very challenging.

A strong alternative to polling, would be text mining and use of speech pattern recognition to predict the results of the 2020 presidential election. Using machine learning algorithms to analyze 100s of tweets and dozens of speeches from past candidates, certain word usage may be a leading indicator to an election victory.

The 2020 presidential election will once again be the most “important election” of our time, but will text mining twitter and speech data from President Trump, Former President Obama, Hillary Clinton, Mitt Romney and Bernie Sanders provide foresight into the 2020 election?

Data and Analysis:

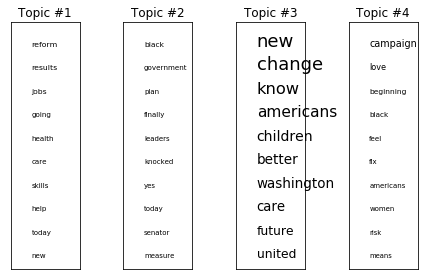
In order to gain a better understanding of the corpora, topic modeling using Latent Dirichlet Allocation (LDA) method was utilized. First stop words were removed specific to the English language, other patters were also removed:

* Speech:
  + All items in parenthesis were removed, most corresponded to audience remarks for speeches, like (boos), (laughter), or (applause). The issue with the speech data was that these were sometimes present and sometimes absent. While a good data point to gather as it provides useful information on candidate’s ability to engage an audience.
* Twitter:
  + Removal of any https (like a shared link) was not useful to the overall strategy
  + Any html code was removed
* Both:
  + Punctuation characters were removed, like apostrophes, commas, semicolons, etc.
  + Numeric terms were removed

Speeches utilized:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Title** | **Person** |  | **Speech** | **Person** |
| Obama Inaugural Address 20th January 2009 | Obama |  | Primary Acceptance speech | Romney |
| Acceptance Speech at the Democratic Convention 2008 | Obama |  | RNC Acceptance Speech | Romney |
| Night Before the Election 2008 | Obama |  | Mitt Romney's Michigan Victory Speech | Romney |
| Final Primary Night 2008 | Obama |  | Mitt Romney’s Speech to The NAACP Convention | Romney |
| North Carolina Primary Night 2008 | Obama |  | Bernie Sanders Iowa 2020 Campaign Rally | Sanders |
| AP Annual Luncheon April 2008 | Obama |  | Bernie Sanders Vision for Democratic Socialism | Sanders |
| Super Tuesday 2008 | Obama |  | Bernie Sanders Speech During Iowa Caucus | Sanders |
| DNC 2007 Speech | Obama |  | Biden Super Tuesday | Biden |
| Concession Speech | Clinton |  | Joe Biden Iowa Speech June 2019 | Biden |
| Speech At Ohio State Summer 2016 | Clinton |  | Joe Biden South Carolina Victory speech | Biden |
| Speech at Michigan 2016 | Clinton |  |  |  |

Obama Topic Modeling:



Hillary Clinton LDA Analysis:

The interesting thing about the Clinton LDA analysis, is that it looks quite similar to Obama’s, for instance in Clinton’s Ohio speech, she used similar words like change, American, new, and jobs. This indicates that term frequency-inverse document frequency (TFIDF) may be the best strategy to pursue for predictions and clustering. The following speeches and LDA analysis yielded the following results:

**ROMNEY Primary Acceptance speech:**

trump like future millions state choice renewable billionaires degree tax

**ROMNEY RNC Acceptance Speech:**

say want change democrats know ready life united years like

**ROMNEY Mitt Romney's Michigan Victory Speech:**

like businesses principles innovation leadership hear available false long person

**ROMNEY Mitt Romneys Speech to The NAACP Convention:**

applause americans new know help years best governor family dad

**SANDERS Bernie Sanders Iowa 2020 Campaign Rally:**

today better campaign life americans promise hope change plan jobs

**SANDERS Bernie Sanders Vision for Democratic Socialism:**

say community wages way hard come pay nation ordinary told

**SANDERS Bernie Sanders Speech During Iowa Caucus:**

going trump want know jobs like americans let new today

**BIDEN Biden Super Tuesday:**

united risk know sacrifice sacred earned pays great prices different

**BIDEN Joe Biden Iowa Speech June 2019:**

research west exist war gay telling try ending wealthy thinking

**BIDEN Joe Biden South Carolina Victory speech:**

want washington years compete economy working ask jobs economic mccain

**BIDEN Joe Biden Town Hall in Davenport, Iowa 2019:**

years white know health care new men end change government

**OBAMA We the People:**

democrat taxes week night spend costs presidential children farmers false

Next in order to vectorize, the following approaches were used in tokenization with unigram and bigram:

* Count Vector: this will just count how often a word occurs in each text
* Normalization of Count Vector: taking the frequency of a word occurrence and dividing by the total number of words in each text
* Inverse Vector: if a term occurs often then the numeric value assigned will be low, if the term occurs often then the numeric value will be high. This puts emphasis on lesser occurring words and phrases
* TFIDF Standardized Normalized/Standard Normalized: this will use the scaler and look at min and max of each word in context of the document and in context of the corpora

The Results of Unigram:

|  |  |  |  |
| --- | --- | --- | --- |
| **Vectorization Strategy:** | **Accuracy Scores** | | |
| GaussianNB | MultinomialNB | BernoulliNB |
| TFIDF Standard Normalized | 42.9% | 85.7% | 100.0% |
| TFIDF Normalized | 57.1% | 57.1% | 42.9% |
| TFIDF Standard Count | 57.1% | 71.4% | 42.9% |
| Count Vector | 42.9% | 42.9% | 42.9% |
| Normalized Count | 42.9% | 28.6% | 42.9% |

While 100% seems high, the limited training and test set (23 speeches total), with the test set at 7. This is very small, however the speech data is an important way candidates communicate with their supporters.

Bigram predicted slightly more accurate scores especially using the multinomial Naïve Bayes:

|  |  |  |  |
| --- | --- | --- | --- |
| **Vectorization Strategy:** | **Bigram Accuracy Scores** | | |
| GaussianNB | MultinomialNB | BernoulliNB |
| TFIDF Standard Normalized | 71.4% | 85.7% | 71.4% |
| TFIDF Normalized | 71.4% | 100.0% | 28.6% |
| TFIDF Standard Count | 85.7% | 28.6% | 28.6% |
| Count Vector | 71.4% | 100.0% | 28.6% |
| Standardized Normal | 28.6% | 85.7% | 28.6% |

***Twitter Data:***

Tweets were sourced manually from the following user accounts and time frames:

* Trump: June – October 2016
* Obama: June – October 2012
* Sanders: February- June 2016
* Clinton: June – October 2016

While Tweepy API is powerful in its ability to access twitter data, it has limitations in sourcing twitter data from any date beyond seven days ago. Thus a stratified sample was incorporated to target important times in an election. About 60 tweets were sourced for each individual during those time frames, and were generally randomized to collect a proper sample set to train a model. From there two strategies were utilized to predict the election: Naïve Bayes and Support Vector Machine Classification.

Both strategies looked to see how Biden’s tweet would classify: Obama, Trump, Sanders, or Clinton. The training vs testing set was split 80%/20% for both strategies.

Naïve Bayes Classification:

The following results show the accuracy of the model using TFIDF Standard Normalization and count vectorization.

|  |  |  |
| --- | --- | --- |
| **Vectorization Strategy:** | **Unigram Accuracy Scores** | |
| GaussianNB | MultinomialNB |
| TFIDF Standard Normalized | 64.9% | 77.2% |
| Count Vector | 70.2% | 75.4% |

Similar to the speech data, the TFIDF standard normalized showed the most promising scores, but count vector seemed to hover around 70-75% for both Gaussian and Multinomial.

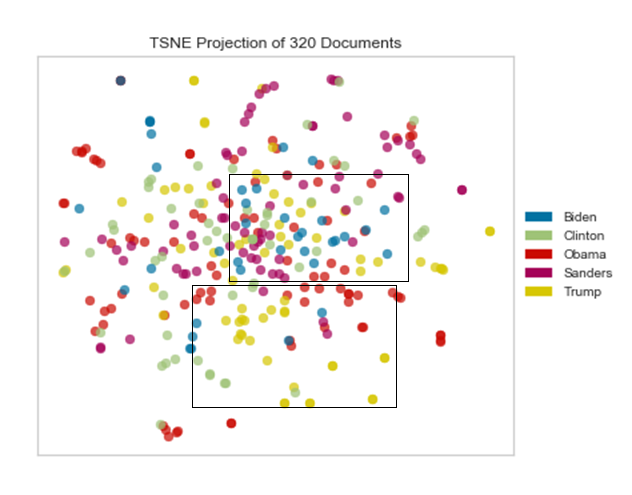
SVM Classification:

|  |  |
| --- | --- |
|  | Accuracy Metric |
| TFIDF Standard Normalization | 66.7% |
| Count vector | 68.4% |

The count vector yielded slightly better results than the TFIDF in this scenario, but both were slightly less accurate at prediction than the NB models.

Clustering:

Using an agglomerative clustering, which clusters using a bottom up approach (i.e. each observation or document starts in its own cluster and clusters are successively merged together) using a distance metric which measures distances between data points.



Using Kmeans to cluster seems to visualize how tweets cluster together seems to indicate that Biden is clustering with Sanders and Clinton on a number of tweets. Also interesting, Obama and Trump seem to cluster together.

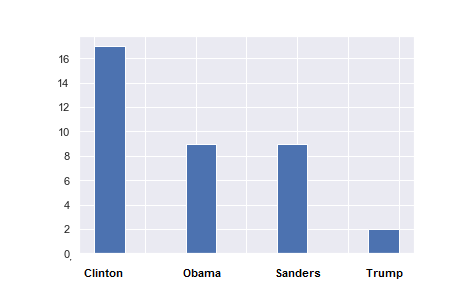
Results

From the Speech data collected, two damning indicators show that Biden generally doesn’t speak like Obama. And it would appear that he speaks more like Clinton, Sanders, or Romney. The only bright spot for Biden is that victory speech in South Carolina may have helped his supporters to

|  |  |  |
| --- | --- | --- |
|  | TFIDF Standard Normalized MultinomialNB | TFIDF Standard Normalized BernoulliNB |
| Accuracy Measure | 85.7% | 100.0% |
| Super Tuesday (March 2020) | 0 | 0 |
| Joe Biden Enters Race (June 2019) | 0 | 0 |
| South Carolina Victory Speech (March 2020) | 1 | 0 |
| Joe Biden Town Hall in Iowa (August 2019) | 1 | 0 |

Twitter Data:

Using the multinomial NB model with an accuracy metric of 77% classification, 37 of Biden’s recent tweets (without reference to covid19) were fed into the model and yielded the following results.

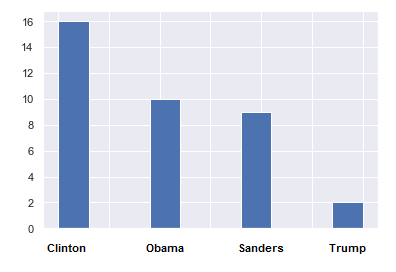


A majority of tweets are classified as Clinton, and then Sanders and Obama. Another telling metric for model performance is that only two tweets classify as Trump, indicating that this model takes into account the difference between Democratic and Republican rhetoric.

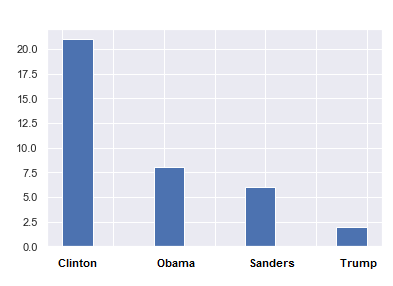
The words most likely to be classified as Clinton:

* 'pence', 'paying', 'pattern', 'pathetic', 'path', 'pass', 'partner', 'per', 'part', and ‘parents’

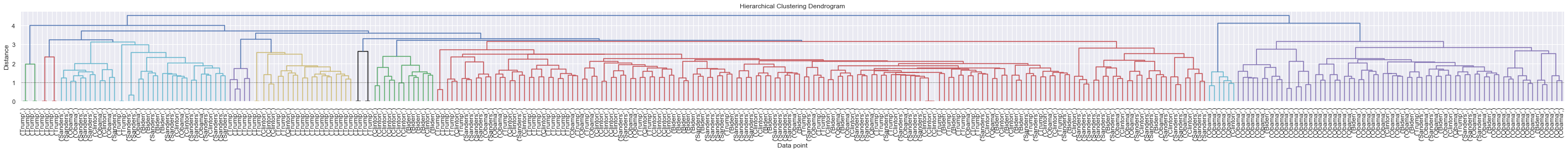
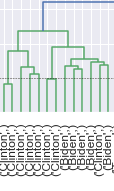
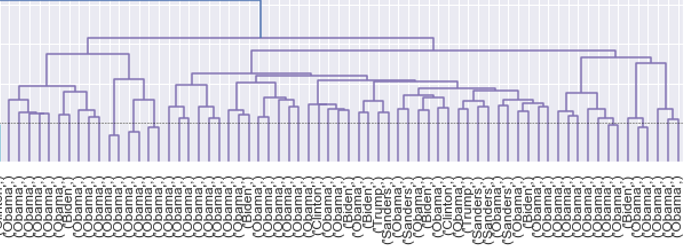
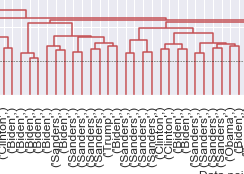
The Count Vector based on multinomial Naïve Bayes shows similar results below.



SVM showed additional skew to Clinton tweets.



Clustering:



Interestingly, a majority of Biden’s tweets clustered together with Clinton and/or Sanders. Trump’s tweets tended to cluster together separately from the pack; as did Obama’s tweets. These results are incredibly concerning for Biden, generally the data tended to lean one way: Biden will have a very tough time getting elected based on past failures from Romney, Sanders, and Clinton.

Limitations in the Results and Mitigates to Limitations

The speech data is a very limited sample, while information can be gleaned from any sample, it is difficult to truly trust the results from a 23 documents. However, generally most of the models did point to Biden’s speech patterns being closer to the rest, than Obama.

Twitter user policy is constantly updated and character limits are especially important when sourcing data. Looking at Obama’s tweets from 2012 show much shorter tweets (less than 140 characters). However looking at overall users, it appears that only 9% of tweets hit twitter’s 140-character limit according to Techcrunch.com. Also, users in today’s ‘twitterverse’ tend to share links and provide opinions or thoughts, whereas in 2012 it was more strictly a quick opinion or thought on a matter.

Finally, the premise that Biden needs to be Obama to win an election is also flawed. Any candidate can create their own persona and charge to an election victory. Also, Biden was born in Pennsylvania a critical swing state and likely has stronger appeal there than Hillary. However, Biden may have roots in Pennsylvania, but that does not necessarily translate to a victory there, especially when the data seems to indicate that Biden’s speech patterns are far closer to candidates that lost there and in the general election.

Conclusion

Based on word usage, Biden will have a very difficult time getting elected in November. As of March 19th, it looks like Biden has all but tied up the Democratic Primary. However, based on speech patterns and vocabulary usage, Biden will struggle to unseat the incumbent Trump.

While the Covid-19 outbreak remains a wild card for the 2020 election, it strongly appears Biden will be a longshot. A lot of the criticism of the 2016 Clinton campaign was lack of excitement to get to the voting booths from the Democratic base. It strongly appears that Biden will suffer the same fate.

Unless Biden can make drastic changes to his campaign now, the slogan “Four more years!” will reign true.