# The Twitter Primary

*The project is designed to visualize what twitter users are talking about for the 2016 presidential election and their preference towards different candidates based on real-time twitter data. All visualizations would be updated every 5 minutes. A tree map is used to visualize the composition and percentage of high frequently mentioned words on twitter. The results of sentiment analysis on tweets were visualized through a scatter plot. The two axes correspond to the two dimensions that we used for sentiment analysis: polarity and objectivity. Also, a bar chart was used to visualize the polarity of tweets particularly.*


| | | |
|---|---|---|
| Linfeng Zhou | lz1335@nyu.edu | lz1335 |
| Yunong Cao | yc2474@nyu.edu | yc2474 |
| Yi Zhang | yz3204@nyu.edu | yz3204 |
| Zeyu Jiang | zj473@nyu.edu | zj473 |


Project page (on Github):
https://github.com/NYU-CS6313-SPRING2016/Group-12-INET-2016-Candidates/tree/gh-pages
Video:
https://vimeo.com/167397229
Working demo:
http://nyu-cs6313-spring2016.github.io/Group-12-INET-2016-Candidates/index.html

**What is the problem you want to solve and who has this problem?**

The result of president election poll demonstrates the perspective of all voters. However, such result is not able to solve all problems since election poll is subject to all voters. What if we are interested in the preference on candidates of just a specific group of people such as Twitter users? In this project, we want to investigate this problem and develop an UI to help those people who are interested in this topic to find out what twitter users are talking about for the 2016 presidential election and their preference towards different candidates.

People who are interested in politics would definitely benefit from our UI to get an intuitive idea about how to use social media to help them get campaign success. In addition, the campaign group of various candidates can also leverage our UI to make and adjust their campaign strategy to attract a certain group of voters.

**What questions do you want to be able to answer with your visualization?**

- Who is the most popular candidate on twitter according to the real-time data?

The number of tweets related to each candidate will be different each time we fetched the data. The percentage composition of tweets will be visualized to find which candidate is talked more comparing to others.

- Which words are discussed most about different candidates on twitter?

People usually talk about different things towards different candidates. What words are mentioned more when people talking about a specific candidate?

- What sentiment are expressed when people use the words mentioned above?

For some words or topic mentioned frequently, whether positive or negative emotion are expressed?

- What are people's attitudes towards each candidate in terms of sentiment?

Based on the sentiment analysis result of tweets, what do people think of each candidate? Two dimensions, polarity and subjectivity, were used to evaluate people's potential attitudes.

- Which candidate is receiving more positive/negative attitude comparing to others?

A value of positive/negative attitude will be assigned to each tweet based on the sentiment analysis, the average value of all tweets related to a candidate will be calculated and compared with others, from which we could know who is getting higher/lower score, thus more likely to be supported/unsupported by twitter users.

**What does your data look like? Where does it come from? What real-world phenomena does it capture?**

Since use Twitter Stream API to fetch real-time tweets, the response data is a txt file in JSON format, which includes extra information about timestamp, user location, geo-tag and etc. in addition to the tweet text.

The attribute we are going to use in this project is listed in the table below:

| Attribute Names | Attribute Type | Meaning |
|---|---|---|
| text | text | The text of tweet that is going to be analyzed |

However, before we finally visualize the information, text is going to be processed first. Specifically, we will perform a sentiment analysis on it. By conducting such analysis, the

subjectivity/objectivity and polarity of each tweet is calculated via Python TextBlob. Two sentiment scores (one positive and one negative) are then be calculated by taking the average of total polarity score that belongs to a specific candidates. For example, if we have 100 tweets with respect to Trump that express negative sentiment with a total polarity score of -70, the positive sentiment score for Trump will be -0.7. In addition, we find out the frequency of each word by counting its occurrence in all tweets after all necessary pre-processing treatment (removing stop words, lemmatizing, stemming) are applied on tweet corpus.
The attributes we visualize on our UI are:

| Attribute Names | Attribute Type | Meaning |
|---|---|---|
| candidate names | text | Names of the candidates |
| keywords | text | Popular words appeared in the tweets |
| keywords popularity | quantitative $[0, \infty)$ | The number of occurrence of the keyword in all tweets |
| subjectivity/ objectivity | quantitative [0,1] | The score that indicates the extent of expressed subjectivity/objectivity in a tweet (0.0 is very objective and 1.0 is very subjective) |
| polarity | quantitative [-1,1] | The score that indicates the extent of expressed positive/negative sentiment in a tweet (-1.0 is very negative and 1.0 is very positive) |
| sentiment score | quantitative [-1,1] | The average of polarity for all tweets against a specific candidate |

**What have others done to solve this or related problems?**
In the past decade, sentiment in Twitter data has been used for prediction or measurement in a variety of domains, including stock market, politics and social movements. A couple of researches focused on the sentiment analysis of twitter data on the US election.
FiveThirtyEight has conducted a project called “The Facebook Primary”, which visualized where 2016 presidential candidates were winning based on the number of likes of verified candidate Facebook. In their project, they counted and compared the number of likes candidates in each county and labeled each county with an unique color representing the leading candidate in that county. The Facebook Primary aim to investigate the attitude of social network users towards

different candidates in a geographical level. Similarly, we also want to conduct sentiment analysis based on the data of social network.

Wang and collaborators have developed a system for real-time analysis of public sentiment toward presidential candidates in the 2012 US election as expressed on Twitter. The system analyzes sentiment in the entire Twitter traffic about the election using naïve Bayes model as the classifier. Through their dashboard, the information of volume and sentiment by candidate, trending words and system statistics were delivered, offering people a timely perspective on the dynamics of the electoral process and public opinion. Similarly, we also focus on the sentiment analysis of real-time twitter data in our project.

**Initial Mockup**

Our initial mockup is showing an interactive map which looks a bit similar to the one made in The Facebook Primary project. The map shows in Figure 1 is a very primitive mockup of our final deliverable.

The main part of the visualization is a map of United States where each state is labeled by a specific color representing the most popular candidate in that state based on real-time sentiment analysis on tweets. (For each candidate, the map assigns a specific color for them. For example, pink stands for Carson and green stands for Trump). The lower left part is a horizontal bar chart showing the real-time support rate of various candidates in the state which users choose. In the lower right corner, there is a list of candidates where users can choose as many of candidates as they want to compare.

For instance, if a user is interested in comparing the real-time support rate of Carson, Sanders and Trump on Twitter, he can select these names in the lower right corner and the map should automatically reflect the leading candidate in each state by showing his/her representing color. If he clicks on a particular state, the lower left bar chart will also tell him a detailed rank of support rate in that state. (If he clicks nothing, the bar chart will show the situation of the entire country.)

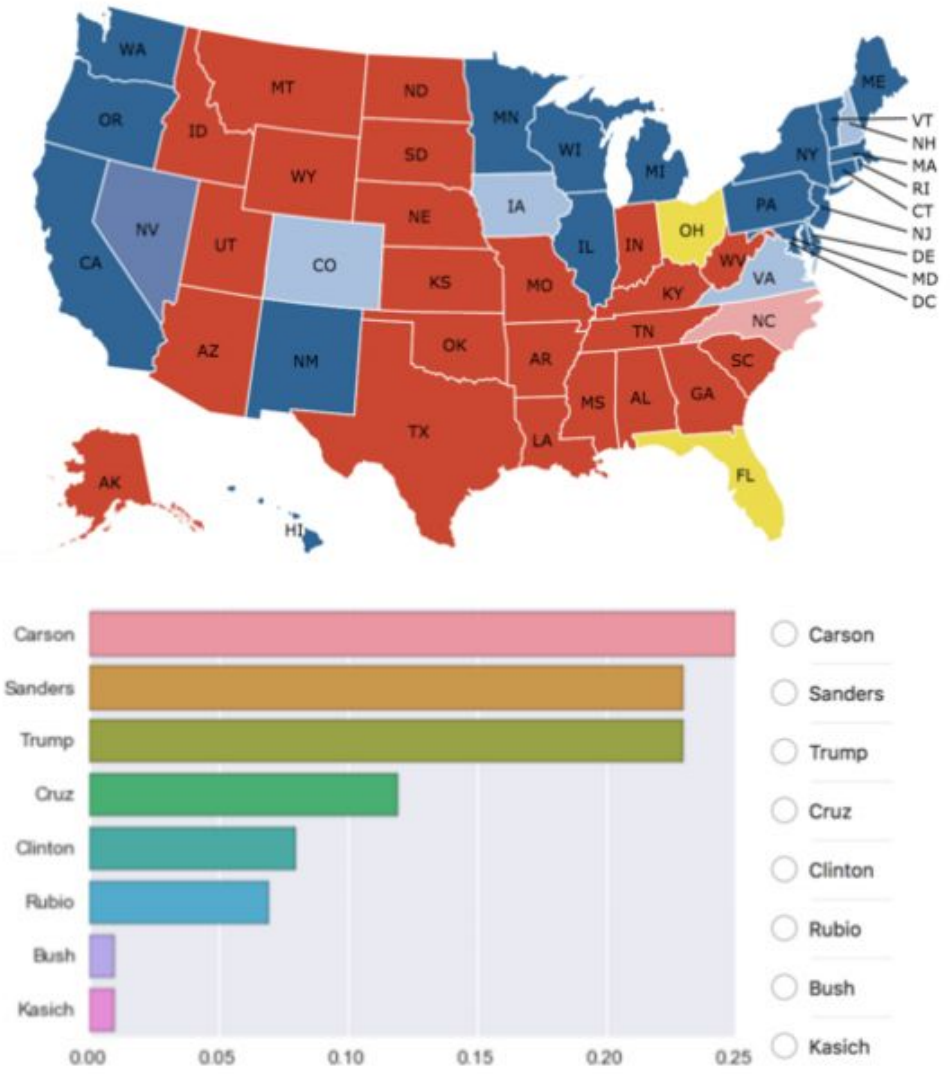


**Figure 1 Visualization Designed in the Proposal**

**Project Update**

By exploring couple of real tweets data, we found there are only few tweets that include geolocation information. Therefore, we doubt whether we have enough data to populate our map by county/census block given that we are using just the real-time data. Another problem of using coordinates (then geolocating them) is that the location when tweets were sent does not necessarily represent the residency of twitter users. Due to all the problems that are mentioned above, we decide to discard the idea of geographical facet. Instead, we want to explore the high

frequency keywords when people talk about different candidates. In addition, we will analyze the sentiment of tweets for each candidate in two dimensions: polarity and subjectivity.

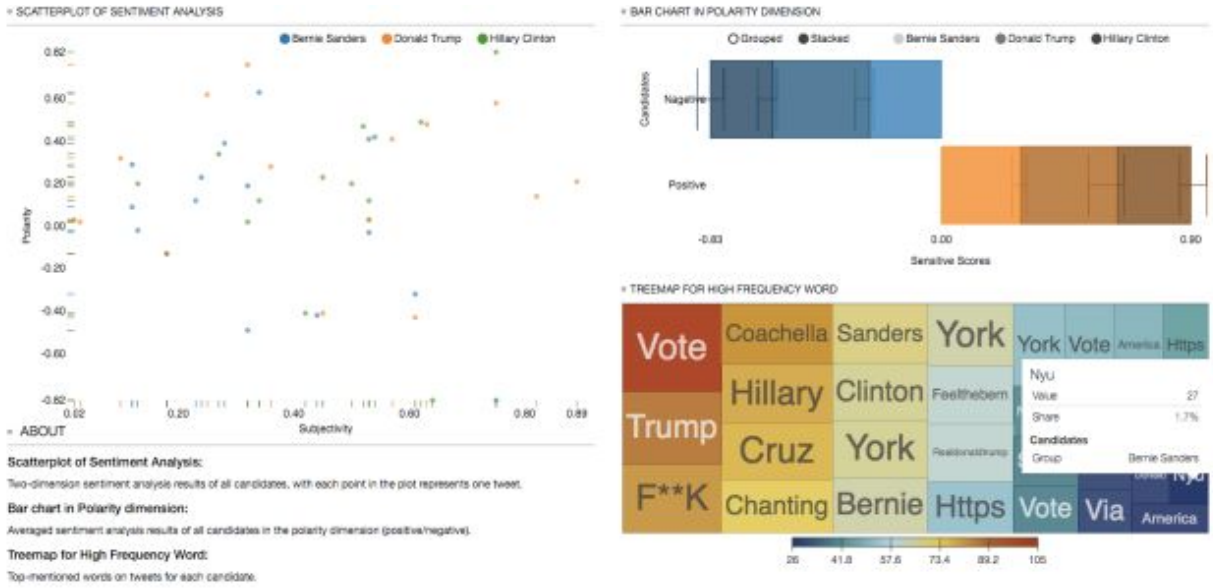


**Figure 2 Overview of the Update Demo**

The two-dimension (polarity and subjectivity) sentiment analysis is displayed through a scatter plot. Each point in the plot represents one tweet, and the color of the point stands for the candidate this tweet is related to (See Figure 3).

A bar chart is used to display the sentiment analysis results for all candidates. Different colors will be assigned to different candidates. For each candidate, the positive sentiment score is the average of polarity for all positive tweets, while the negative sentiment score is the average of polarity for all negative tweets. Three implementations are used to view the sentiment analysis results in different aspects (See Figure 4):

1) Candidate Name

By selecting the name of a candidate, the sentiment analysis results related to him/her will be displayed.

2) Group

Under the group selection, the results of one or more selected candidates could be compared.

3) Stack

The stack selection will display the magnitude of the positive score over all candidates versus the magnitude of negative score over all candidates. This implement is aimed to get a general understanding of the overall positive/negative attitude of twitter users in the collected data.

(Note: If there is only one candidate selected, the visualization of group and stack will be the same.)

Tree map is used to demonstrate which words are discussed more on tweets related to each candidate. The more frequent a word is mentioned, the more area will be assigned to it. High frequency words are represented by warm colors while low frequency words are represented by cool colors. When hovering on a specific word, the related candidate could be shown under "group" inside the tooltip (See Figure 5).

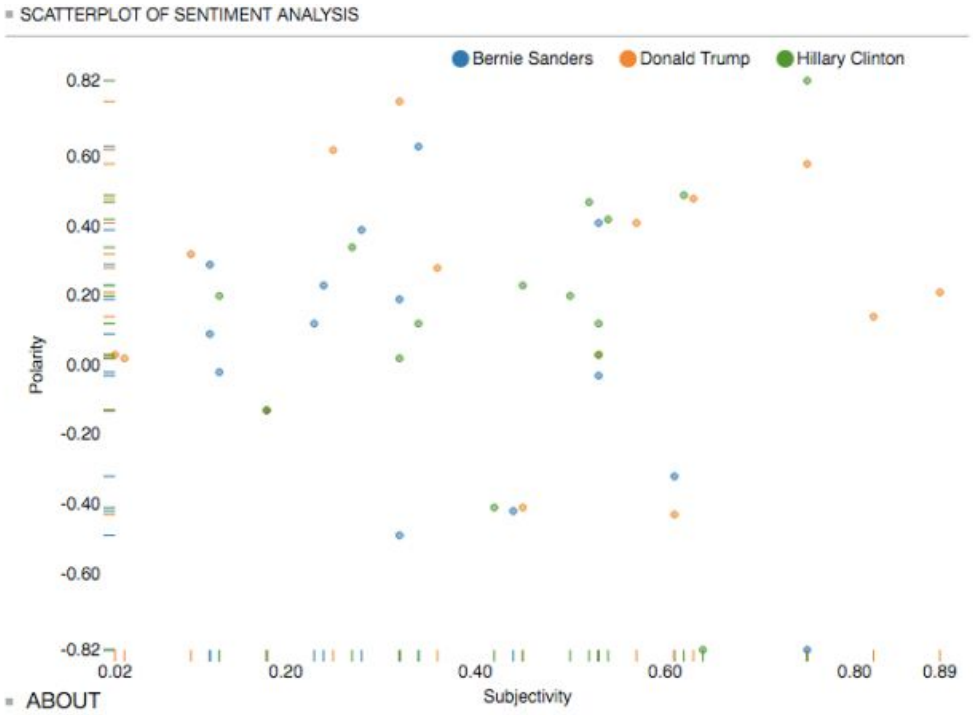


**Figure 3 Scatter Plot of Sentiment Analysis**

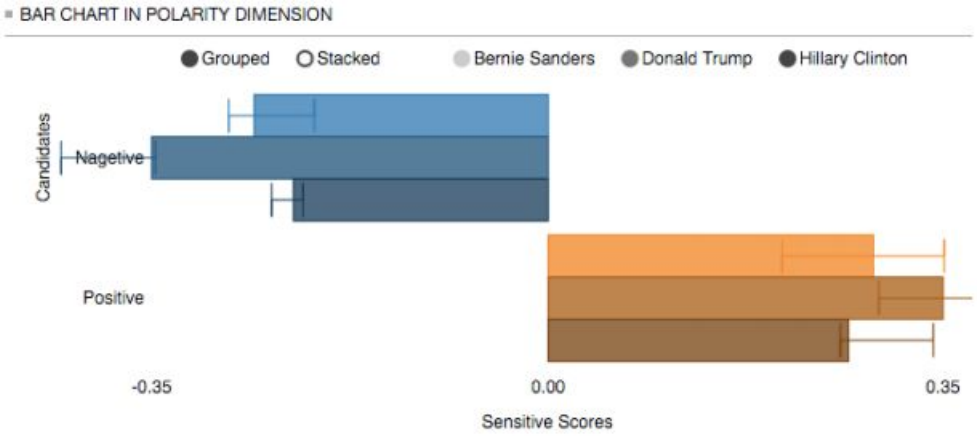


**Figure 4 Bar Chart of Sentiment Analysis in Polarity Dimension (with All Candidates Selected)**

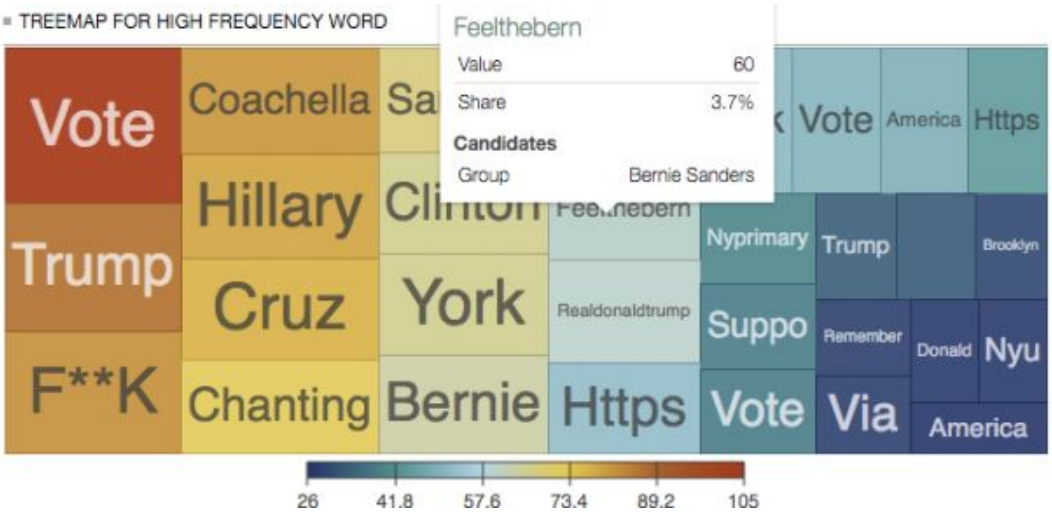


**Figure 5 Tree Map for High Frequency Word When Hovering on a Specific Word ( in this case, Feelthebern)**

## Final Visualization

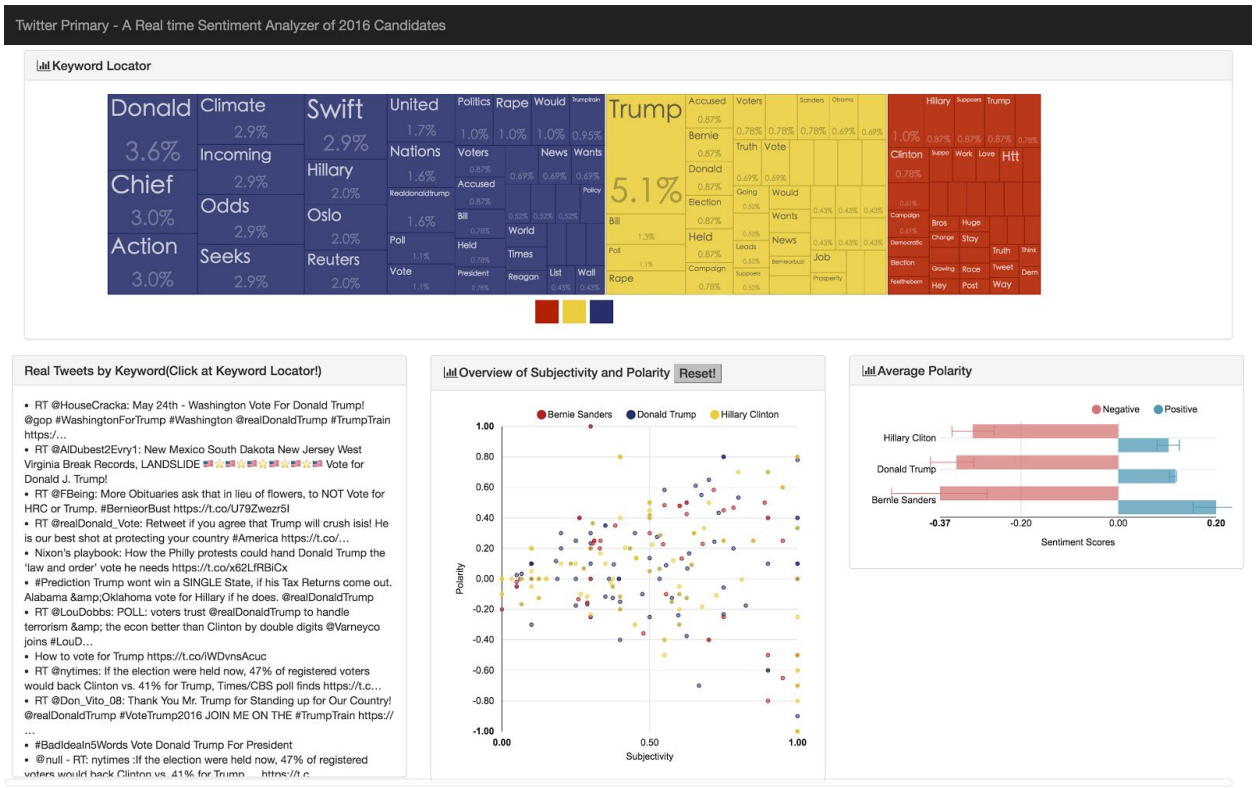


**Figure 6 Overview of the Final Visualization**

We reorganize the structure of dashboard and make the Keyword Locator as our main visualization. In order to find potential patterns in keywords when people are talking about

different candidates, we create three individual tree maps for each candidate instead of mixing all keywords together. It is easy to visualize the composition and percentage of high frequent mentioned words for each candidate. The area of a keyword is positively correlated to its frequency mentioned in tweets. The color stands for the candidate that the keyword is related to: red for Bernie Sanders, yellow for Hillary Clinton, and blue for Donald Trump. Detailed information of each word (value, share, related candidate) could be viewed in a tooltip when hovering on it.

In the bottom left, we add a panel to show the actual tweets. Each time the user is clicking on a keyword, all related tweets would be displayed in the Real Tweets by Keyword.

Next to the Real Tweets section is a scatter plot showing the results of the two-dimension sentiment analysis (polarity and subjectivity) for each tweet. Each point in the scatter plot represents one tweet, with the sentiment analysis results as the coordinates: x-axis value for objectivity (from 0.0 to 1.0) and y-axis value for polarity (from -1.0 to 1.0). Similar as above, the color of each point stands for the candidate that this tweet corresponds to. By selecting the candidate name on the top right, only the tweets related to the selected candidate will be displayed. When hovering on a point, the corresponding tweet would be showed in a tooltip. The scatter plot is also interactive with the Keyword Locator: when a keyword is selected, only the sentiment value of the tweets that contain this keyword will display in the scatter plot. By clicking the “Reset” button, the scatter plot of all tweets will be resumed.

We also visualized the polarity of tweets particularly in the Average Polarity section. We discard the idea of stacked bar chart, and decide to show the sentiment score in positive and negative polarity separately. For each candidate, the positive sentiment score is the average of polarity for all positive tweets related to that candidate, while the negative sentiment score is the average of polarity for all negative tweets. By selecting positive/negative on the top right, the positive/negative sentiment scores for all candidates could be compared separately.

In addition, we add a countdown timer at the top of the page to make users aware of the next refresh time. All visualizations are updated every 5 minutes according to the real-time tweets. The page could also be refreshed manually by clicking on the “Next Refresh” button.

**Data Analysis**

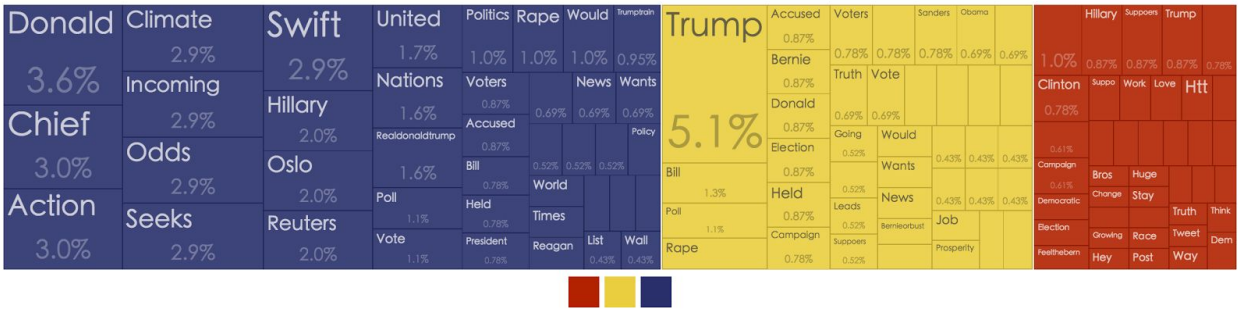


**Figure 7 Tree Map for High Frequency Word for All Candidates**

**For an overview:**

- **Who is the most popular candidate on twitter according to the real-time data?**

Figure 7 shows a close look of the keywords locator, the main component of our UI. At a glance, one can notice that first, there are much more people talking about Donald Trump than the other candidates, which can be seen by noticing the fact that the area covered by blue color is larger than those covered by either yellow or red (The exact number of tweets belongs to a specific candidate can be shown by docking the mouse on the three little squares at the bottom of the keywords locator). This is as expected since Donald Trump is probably the most talked-about people throughout the world now.

- **Which words are discussed most about different candidates on twitter?**

Second, the chart tells us that people like to use words "Chief", "Action", "Climate", "Swift" and etc. when talking about Trump as we are writing this and we are going to find out why people address these words later. Third, by noticing the largest area in the yellow block, we can see that people like to compare with Donald Trump when talking about Hillary Clinton. We also found people like to mention her husband and talk about the rape case on Twitter since words "Bill" and "Rape" appear in the keywords. Fourth, for Bernie Sanders, people like to use "Support", "Work" and "Love" when talking about him.

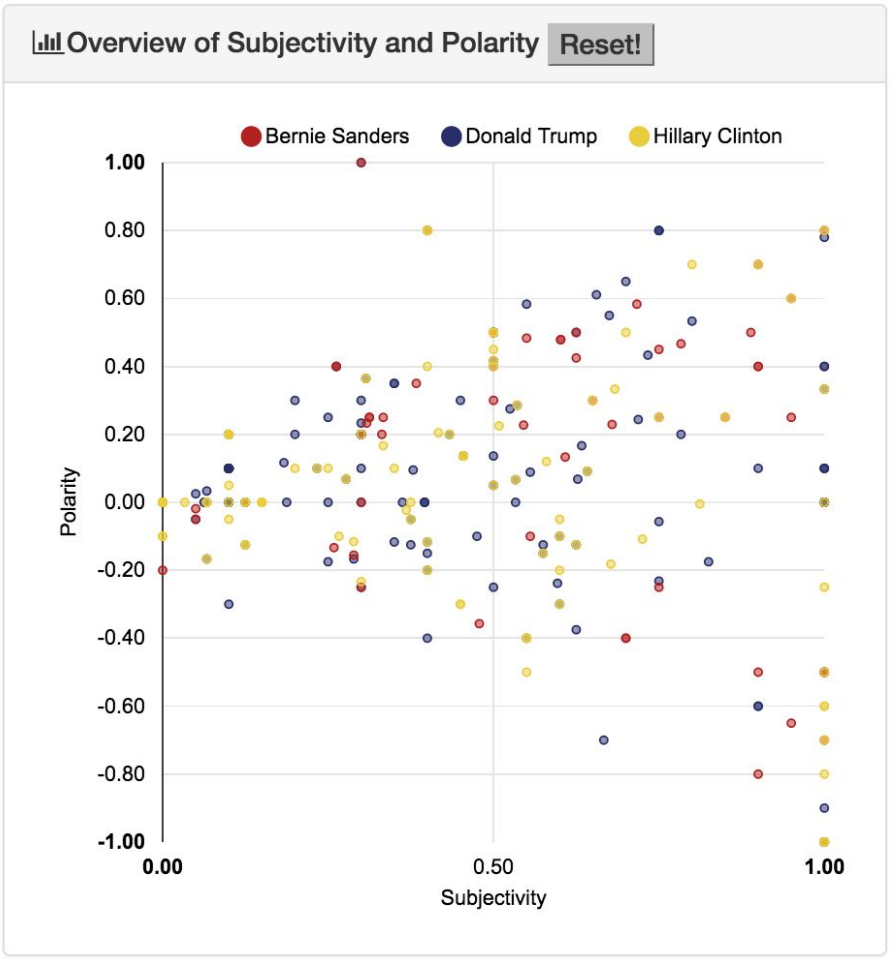


**Figure 8 Scatter Plot of Sentiment Analysis (with All Candidates Selected)**

- **What are people's attitudes towards each candidate in terms of sentiment?**

Figure 8 shows an distribution of both polarity (on y-axis) and subjectivity (on x-axis) as a result of sentiment analysis on all tweets. Generally, the distribution of tweets belonging to three candidates follow the same pattern: first, a majority of tweets are clustered in the low polarity (close to 0) region. Second, the more positive or negative sentiment expressed in tweets, the more subjective they tend to be. Another interesting finding is that, we see a lot more subjective tweets with respect to Donald Trump and Hillary Clinton but not that many for Bernie Sanders.

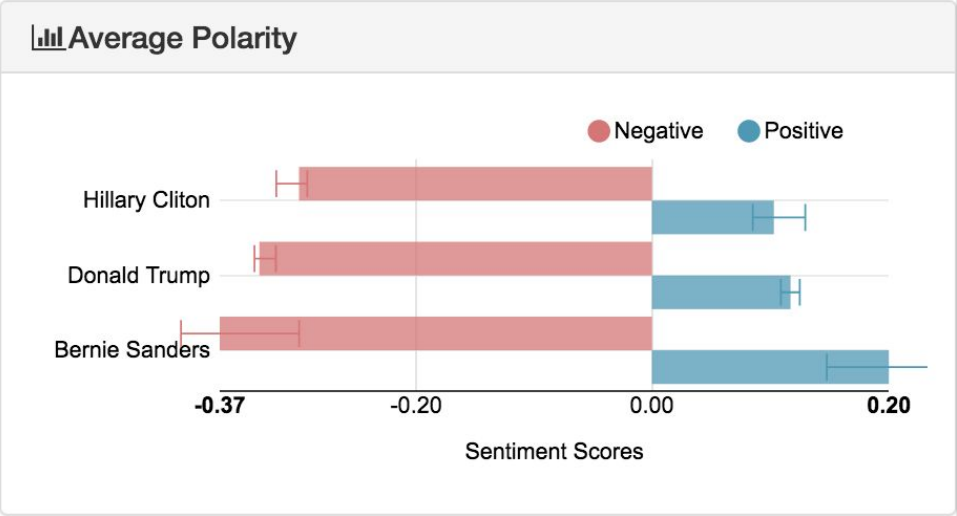


**Figure 9 Bar Chart of Sentiment Analysis in Polarity Dimension (with All Candidates Selected)**

- **Which candidate is receiving more positive/negative attitude comparing to others?**

By comparing the average polarity as shown in Figure 9, we find people express more negative emotion than positive one on Twitter when talking about the three candidates. In addition, we can see that people tend to have both more positive and negative sentiment against Bernie Sanders on Twitter. In other words, more polarization is seen for Bernie Sanders than the other candidates. However, one should also noticed that the error (the marks on the end of bars) is also larger for Bernie Sanders since there are far less tweets belongs to Bernie Sanders in comparison with the other two.

**Case study:**

- **What sentiment are expressed when people uses some keywords?**

We just noticed figure 10 tells us that people like to use words "Chief", "Action", "Climate", "Swift", “Oslo” and etc. when talking about Trump as we are writing this and we are interested in figuring out why and what sentiment are expressed when people using these words. By clicking those keywords in the keywords locator, the real tweets related to these keywords emerges in the lower bottom corner. It turns out many people are forwarding the news "Incoming U.N. climate chief seeks swift action: at odds with Trump" on twitter when we are writing this. In order to check out whether positive or negative sentiment is expressed when people tweeting such words, one can take a close look of the scatter plot. The scatter plot now shows the sentiment distribution only for the tweets selected in the keywords locator. (as shown in figure 11). It seems that tweets with respect to these words are generally subjective (with subjectivity score of 0.1 to 0.3) and positive (with polarity score of 0.1 to 0.3). To check what exact tweet/tweets leads to a particular sentiment score, we dock our mouse on the specific data point and the tool tweet shows that tweet is “Action/ paid $ to run ads to stop Trump…”.

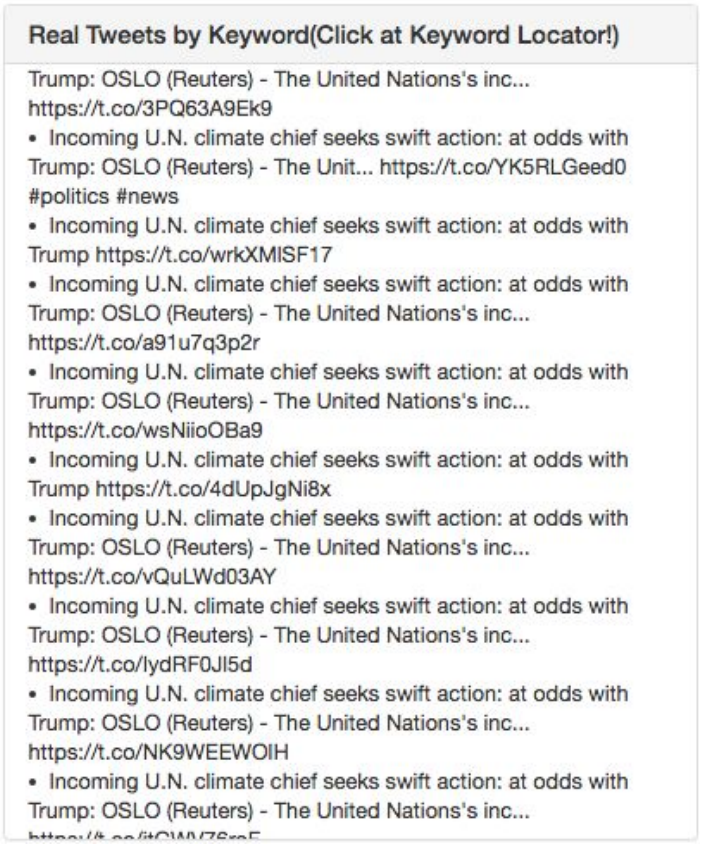


**Figure 10 Real Tweets by Keywords**

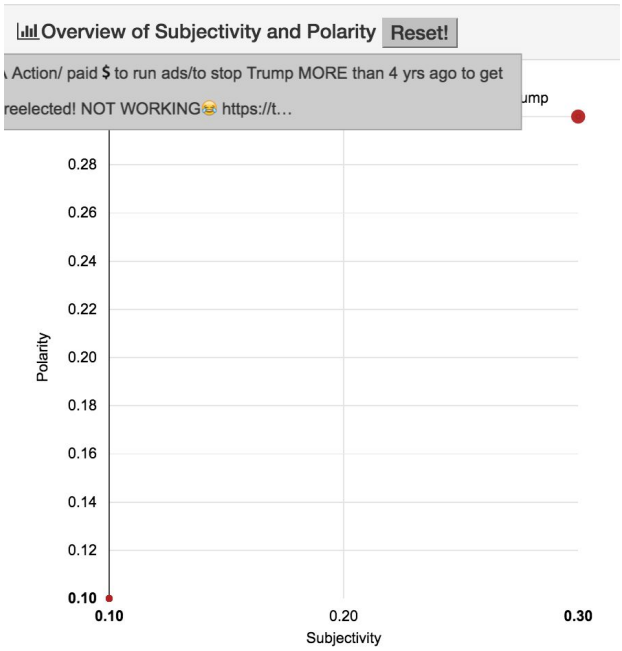


**Figure 11 Sentiment Distribution for Tweets with respect to keywords “Climate”**

**Limitations and Future Works**

Our UI works fairly according to some users we invite. Nevertheless, it still has a lot of space for improvement. First, one major limitation of our project at this stage is that we are not successful to highlight a keyword in the Real Tweets panel when it is selected in the treemap. We will try to address this problem in the future work. Second, the bar chart now has not now interacted with the other two charts. In the future, we want to integrate them so that the bar chart can show an average score of some specific tweets with respect to a particular keyword. Third, the UI now just shows a real-time situation of sentiment on Twitter. Therefore, if funding is available, we could store the historical twitter data in a database and visualize the fluctuation of sentiment score with time, which might provide more useful information concerning the attitude of users towards candidates. Last but not least, the algorithm for sentiment analysis is not accurate enough now. This can be improved by using historical data to perform machine learning and use the result of machine learning to better classify the sentiment expressed in tweets.

**Reference**

Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (n.d.). A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle. Retrieved May 20, 2016, from http://www.aclweb.org/anthology/P12-3020

The Facebook Primary. (2016). Retrieved May 20, 2016, from http://projects.fivethirtyeight.com/facebook-primary/