Quantitative Analysis of Mainstream Media Bias During 2016 Presidential Election Using Large Scale Social Media as Reference

---

Team members: Fangyuan Huang, Tang Xuan

CSC 440 – Final Project Report

2016 Fall

Motivation and background

During the presidential campaign 2016, a question has been repeatedly mentioned, discussed, and debated, yet no clear or solid conclusion has been made: is the mainstream news media at U.S. biased and lost their objectiveness?

Figure 1. Democratic nominee Hillary Clinton (left) and Republican nominee Donald Trump (right)

With the help of rapid technological advancement, news media now has been more reachable then ever and has become one of the main information sources for a great portion of population. As a study by PewReserachCenter shows, in 2016, 57% of U.S. adults often get information from TV, and 38% of U.S. adults get from online websites or social medias. [1]

As a result, news reports from medias could greatly help, or hurt, a presidential hopeful's chances of winning the party's nomination, and further shape the result of presidential election. However, according to a survey conducted by the First Amendment Center and USA Today on November 7, 2015, more than 70 percent of Americans believe that news media is intentionally biased when reporting news, a 17% drop from year 2014. Older audiences are more likely to buy into the media's mantle of objectivity, with 26% of those 50 or older agreeing with the claim. Only 7% of 18-29 year olds agree. Democrats (36%) are much more likely to believe in the news media try to report without bias as opposed to Republicans (19%). [2] This topic has exceptional importance, because the legitimacy of a democratic government largely relies on fair and informed election votes. Citizens cannot cast informed votes or make knowledgeable decisions on matters of public policy if the information on which they depend is intentionally distorted. [3]

In this project, we applied data mining techniques and tried to offer a quantitative analysis to this question: does there exist a noticeable bias among mainstream news medias, and if exists, to what extend? and if doesn't exist, why a certain amount of people are feeling like so? The team focused on the two most disputed candidates, democratic nominee Hillary Clinton and republic nominee Donald Trump to simplify the problem (figure 1).

Before we unfold our analysis, we must give a clear definition about the object we are measuring: media bias. As mentioned in related work conducted by D'Alessio and M. Allen[4], types of media bias we considered here were gatekeeping bias, which is the preference for selecting stories from one party or the other; coverage bias, which considers the relative amounts of coverage each party receives; and statement bias, which focuses on the favorability of coverage toward one party or the other.

The team planned to measure gatekeeping bias and coverage bias by comparing the number of reports related to each candidate on each topic; and to measure statement bias by comparing sentiment of reports between each media and comparing to twitter and other references.

Methodology and Assumptions

The methodology applied in this study aims to sentimentally and quantitatively analysis articles from a wide range of mainstream news media about 2016 presidential election to determine the existence of bias among those investigated medias. Thus, the team essentially wished to take string type input as articles and media reports, convert into numerical output as sentimental scores to draw meaningful insights. To achieve this goal, the team applied following steps:

1) Selecting news media:
   In order to present a diverse and representative sample data, the team picked news media according to two criteria:
   1. Ideology. Team wished to cover media ranging from liberal to neutral and conservative. To determine the ideology of each media, the team referred to an article posted by Washington post on October 21, 2014, "Ranking the media from liberal to conservative based on their audiences" [5].
   2. Audience size and massiveness. Other than the ideology, massiveness of a media is also considered. Since the team was more concerned with mainstream media, small or local medias were thus not proper target for this project.

   As a result, the team chose 5 news media to continue investigation:
   Liberal Media: New York Times (NYT), Washington Post (WP),
   Conservative Media: Fox News (FOX)
   Neutral Media: CNN, NBC

2) Collecting data: A python code with urllib2 and request library was applied to crawl articles related to 2016 presidential election from each selected news medias. All articles that mentioned 'trump', 'donald', 'clinton', or 'hillary' were included. However, since the search function from each media differs greatly and usually is very unfriendly to data crawling, the team didn't search archived articles directly from website of those selected

media. Instead, team utilized search function of google to do the job. For example, to find articles related to Donald Trump, team formed a python code to apply "Donald trump, site: www.foxnews.com" to google search engine, and record returned urls from google search results.
Time constrain: Hillary Clinton declared her presidential announcement on April 12 2015, and Donald Trump declared his presidential announcement on June 16 2015. We took

min(04212015, 06162015) = 04212015 = April 12 2015

to be our time constrain. That is saying, no article posted earlier than April 12 2015 was considered.

During data collection, the team made following two assumptions to simplify the process:

- In one article, the main character must appear more frequently than other less-important characters.
- One article could only be either positive, neutral, or negative (i.e. those attitudes are orthogonal to each other, one article could be positive and neutral at the same time.)

According to the first assumption, team categorized articles to three categories: reporting Donald Trump, reporting Hillary Clinton, or reporting both of them.
Regarding to the second assumption, team restricted sentiment of analyzed articles to three dimensions: positive, neutral, and negative.

3) Sentimental analysis and statistics:
Sentimental analysis is the process of determining the emotional tone behind a series of words, used to gain an understanding of the attitudes, opinions and emotions.[6]
Sentimental analysis takes strings like sentence, paragraph, and article as input, and gives numerical value between -1 and 1 as results, which allow us to perform statistical techniques to draw more insights.
In python, several sentimental libraries were available. The team particularly compared four packages: nltk vader, nltk textblob, IMB Watson API and text-processing API. In the 'data and results' section we would talk into details about these packages and explain why or why not we operated with each of them.

4) Comparison the sentimental results with other sides:
Sentiment analysis alone could not tell us much about bias. For example, a media may has 80% of its news report be about one candidate without making gatekeeping bias, if that candidates happen to have more stories and do more stuffs while another candidate happen to be much more quiet and introspective.
Consequently, in order to determine existence of bias, one would need other side of data as comparison such as social media. Like last example we made in last paragraph, let's

consider a media that has 80% of its news report talking about positive things about one candidate, while at the mean time, more than 60% of twitter users or facebook users are talking negatively about the same candidate. Then it's very likely that there exist bias of that particular media.

The Data and Results

Collected Data of news reports recorded from each selected media:

News reports recorded	Hillary Clinton	Donald Trump
CNN	648	644
NBC	628	664
NYT	620	550
W P	578	579
FOX	782	752
Average	651.2	637.8

Table 1. number of urls collected for each nominee

By running crawling code, the team collected 3256 news urls about Clinton and 3189 news urls about Trump. The difference is within 2.4% (table 1). However, if we consider that Hillary Clinton declared her presidential announcement 2 months before Donald Trump did. She actually got less exposure from media, compared to Trump. This makes sense because Hillary has said by herself that she prefer to do thing in a low-key, and Donald Trump is more like a TV-star that loves spotlight and media exposure.

Sentiment package:
Among nltk vader, nltk textblob, IMB Watson API and text-processing API, the team finally picked vader and texblob to conduct our experiment.

The main reason is the limitation of the other two. Both Watson API and Text-processing API was discarded because it limited its capability that each IP could send at most send 1000 calls per day, other than this amount need extra payment per call to proceed. As a result, team chose two free sentiment tools, vader and textblob to conduct our project. However we must note here, the free tool also introduce more sentiment errors into the investigation.

For example, an article posted by Michael E.Miller and Fred Barbash about Donald Trump's disrespectful speech to John McCain[7] was labeled positive to trump by textblob, which is obviously an error. Vader, on the other hand, returned a negative value for this article as we expected.

However, that's not saying Vader is doing a perfect job. In 'Not our President: Protests Spread After Donald Trump's Election' posted Nov.9 2016 on New York Times, the author talked about massive violent, and angry actions after Donald Trump's election and showed clear negative attitude towards Donald Trump but Vader returned positive value and textblob returned negative value as we expected.

In 'Donald Trump's Taped Comments About Women' posted Oct.8 2016 on New York Times, the article wrote about the notorious 'lock-room talk' of Donald Trump, but both vader and textblob gave positive return simply because more positive words appeared.

From these examples, we could see that both Vader and Textblob are far from perfect. We will discuss more improvement in next section.

Sentiment Results:
Following are the results of vader and textblob analysis for each selected media, with Hillary Clinton in blue and Trump in red (figure 2-6).

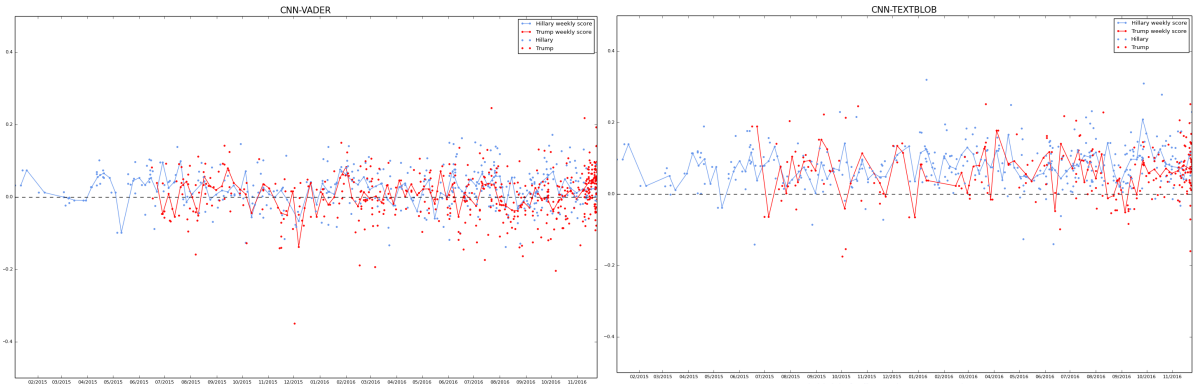

Figure 2. Vader (left) and Textblob (right) analysis of CNN

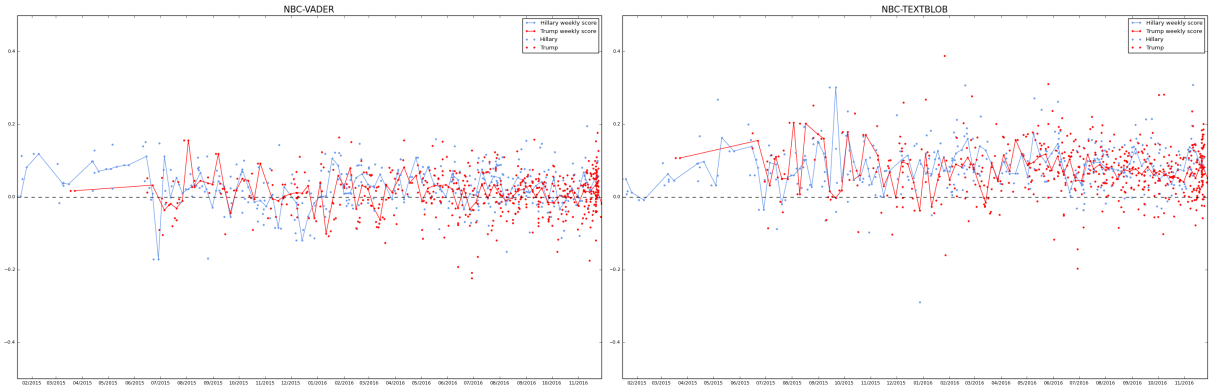

Figure 3. Vader(left) and Textbolb (right) analysis of NBC

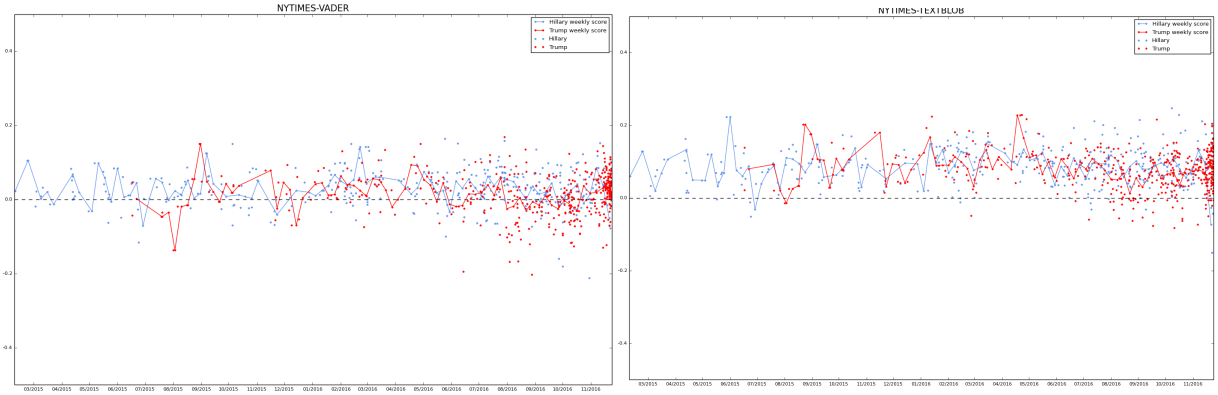

Figure 4. Vader(left) and Textbolb (right) analysis of New York Times

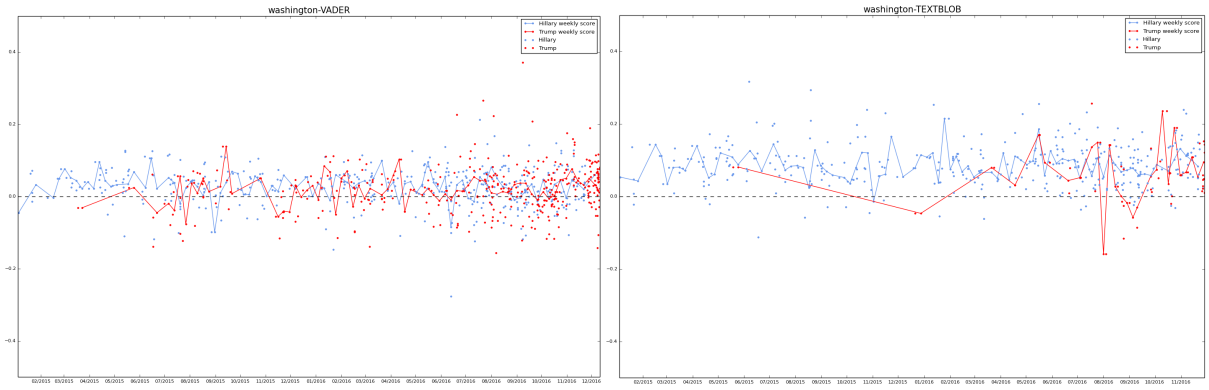

Figure 5. Vader(left) and Textbolb (right) analysis of Washington Post

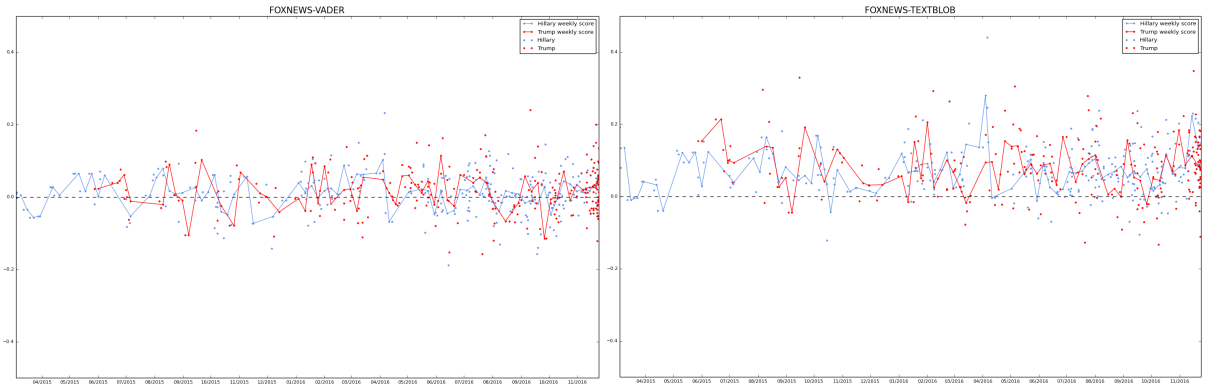

Figure 6. Vader(left) and Textbolb (right) analysis of FOXNEWS

Statistics:
Mean value = μ, standard deviation = σ, and ratio between negative post and positive post =r, were calculated. Higher r value indicates more negative reports. Highest value among each statistics were highlighted in red, and lowest were highlighted in green:

Vader Results	Hillary Clinton	Donald Trump
	Mean, standard deviation, r	Mean, standard deviation, r
CNN	0.025872, 0.056532, 0.463710	0.007334, 0.065602, 0.792453
NBC	0.030028, 0.060750, 0.451163	0.011011, 0.058234, 0.638436
NYT	0.029184, 0.053904, 0.337607	0.013822, 0.056903, 0.530466
WP	0.027084, 0.055703, 0.354478	0.019278, 0.065624, 0.575221
FOX	0.010099, 0.059311, 0.645833	0.018110, 0.064860, 0.648855
Average:	0.024453, 0.057240, 0.450558	0.013911, 0.062245, 0.637086

Table 2. statistic summary of vader's analysis results

From analysis given by Vader, we could see that NBC talked most positively to Hillary Clinton, and New York Times reported Hillary Clinton with highest consistency, and least ratio of negative reports. At the meantime, Fox news reported least positively to Clinton and reported negative news for Clinton with highest ratio.

On the other side, CNN reported least positively for Donald Trump and reported negative news for Trump with the highest ratio. New York Time reported Donald Trump with least negative ratio and highest consistency. Washington post, surprisingly, turned out to be most supportive media to Donald Trump, and New York Times report least ratio of negative articles.
In general, Clinton received consistently higher score than Trump from most selected medias other than fox news. Clinton received more than 100% sentiment scores from CNN, NBC and NYT, which is fairly consistent to the popular impression. And Trump consistently got a lot of negative exposure among all five selected media.

Textblob Results	Hillary Clinton	Donald Trump
	Mean, standard deviation, r	Mean, standard deviation, r
CNN	0.082709, 0.063746, 0.056075	0.064064, 0.074009, 0.201183
NBC	0.087291, 0.067776, 0.072165	0.076157, 0.073925, 0.153670
NYT	0.086355, 0.053657, 0.042071	0.075825, 0.053996, 0.101781
WP	0.086298, 0.061777, 0.070968	0.057140, 0.080617, 0.225000
FOX	0.073046, 0.069967, 0.134615	0.084311, 0.083232, 0.190217
Average	0.083140, 0.063385, 0.075178	0.071450, 0.073155, 0.174370

Table 3. statistic summary of textblob's analysis results

From analysis given by Textblob, the results were more mild, which is consistent to textblob's tendency to count articles as neural. However, the stats still shows that Hillary generally received higher sentiment score and less negative exposure than Trump. Similar to vader analysis, New York Times turned out to be the media that's least favor to posting negative reports.

However, in textblob analysis, Washington post turned out to be the media that's least positive to Donald Trump and report most negative information about him, which totally contradicted to vader analysis. This interesting contradiction revealed part of the difficulty of sentiment analysis, and will be discussed in next section.

Fox News in textblob analysis, similar to vader analysis, turned out to be the media that's least positive to Clinton and comparably most positive to Trump.

Conclusion and Discussion

-----Existence of bias? likely, but we can't say yet.

Despite the ambiguity nature of English and other human languages, team gained some useful information from the statistics about the mainstream media.

Generally, both two sentiment tools showed that Clinton was more favored among mainstream media, especially among liberal and neutral media (CNN, NBC, New York Times, and Washington Post). Only Fox News tended to report Trump more positively than Clinton. All 5 selected media tended to report more negative news about Trump than Clinton (ranging from 41.3%-266%). Conservative Fox news, even if it's the most positive media to Trump and the least positive media to Clinton, still had its negative report vs positive report ratio on Trump high than average.

Due to the late logistic of the raw twitter data, the team didn't manage to run massive sentiment analysis on twitter to reflect twitter users' sentiment about each presidential by our own.

However, the team found a related work which could reveal partial information that could be helpful.

In the ‘Candidates' Speech Analysis Project’ conducted by Erica Kim [8], Kim made the following studyies. The first figure (figure 7.) indicated that both candidates twittered a lot, but Trump received more number of favorites than Clinton.

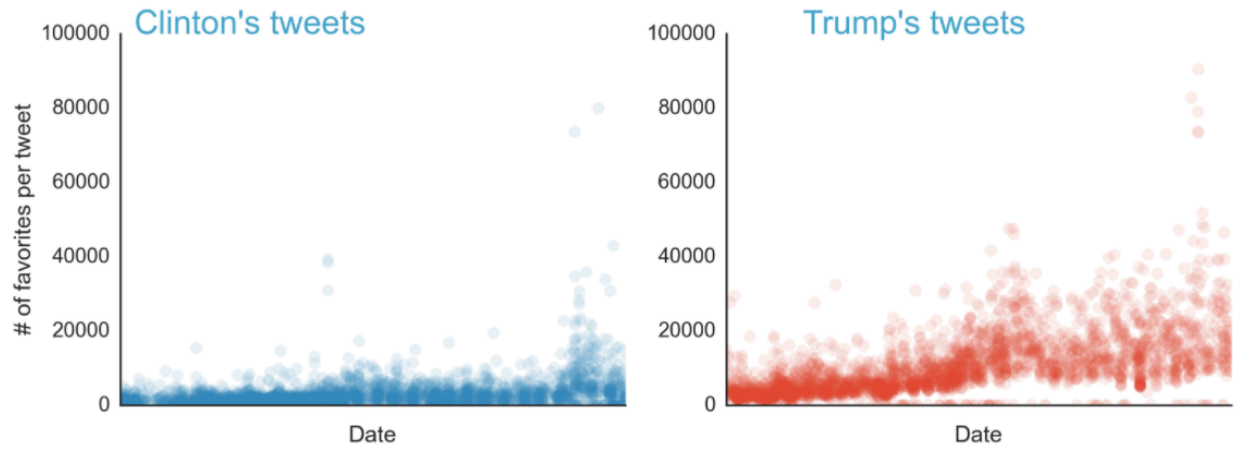

Figure 7. number of favorite per tweet of Clinton and Trump’s tweets according to date time, source: Candidates’ Speech Analysis, Erica Kim

Kim also offered a deeper look into the data and categorized number of favorite of each tweet according to different topics (figure 8). From which, the team found clear tendency that Trump’s tweets received more ‘favorite’ from every topic except higher education. In the higher education topic, Clinton’s tweets got about as many favorites as than Trump’s tweets.

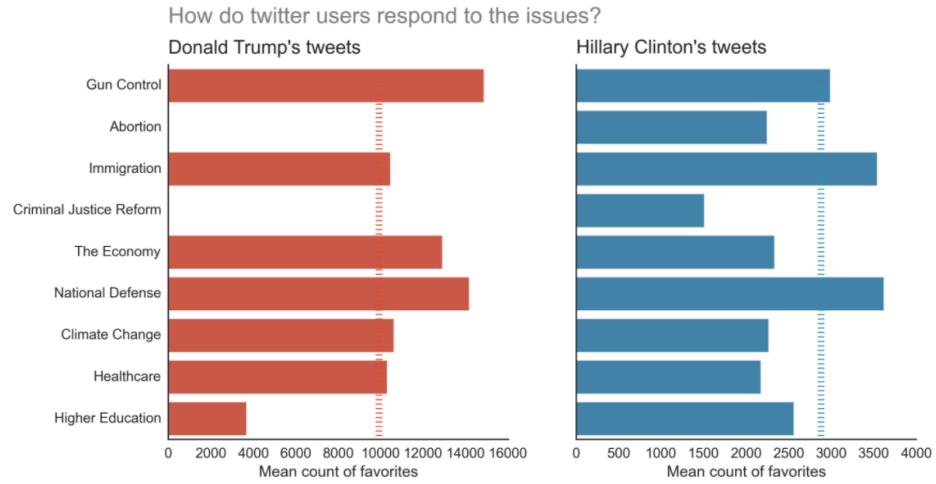

Figure 8. Mean count of favorites of both presidential according to topics, source: Candidates’ Speech Analysis, Erica Kim

In conclusion, combining Mainstream Media data, Tweet data, and final poll result together, the team found an existing and comparatively obvious positive tendency toward Donald Trump on

Tweeter, and an existing relatively strong negative tendency towards Donald Trump on Mainstream Media. The overwhelming preference to Trump on Tweeter might be caused by the nature that Twitter is an international community, and it is possible that some supporter of Trump lives outside of U.S. The negative tendency of media against Trump might be contributed by Trump's preference to spotlight, personal drama style and lack of political experience. Given the population result turned out to be quite close (figure 9), and Donald Trump received more electoral votes, the team agreed that it is likely that there exists a bias against Donald Trump among mainstream media. However, we would need more evidence to make this possibility a solid claim. The reason and improvement would be discussed in the following section.

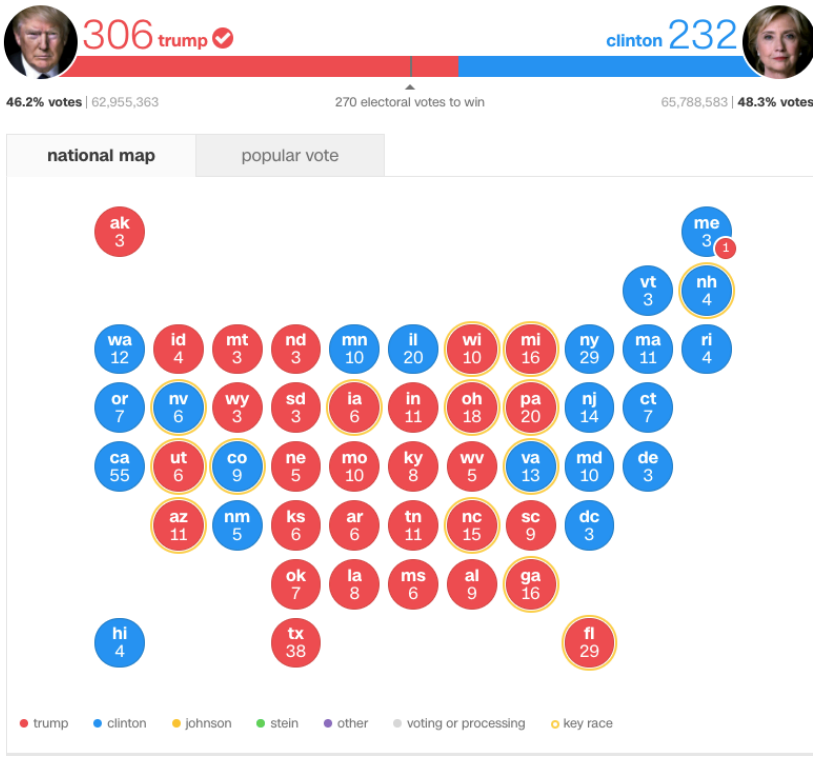

Figure 9. final result of presidential election 2016, source: CNN.com

Limitation and future work

First, Sentimental limitation -- From the discussion at the beginning of 'the data and results' section, it's clear that the sentiment analysis tools we applied was a big limitation to the accuracy of this project. This problem could be greatly improved by more expensive tools. The team planned to carry this project on and use Watson API to conduct sentiment analysis instead of vader or textblob. This should improve sentiment accuracy by a huge amount. What's more, the Watson API could offer sophisticated analysis on keywords and topics, which would be very useful to determine gatekeeping and coverage bias.

Second, Twitter limitation– The absence of twitter analysis is another limitation of this project. Without it, the team wasn't able to compare the sentiment on same topics between twitter users and mainstream media according to our definition on media bias (see in motivation section). The team received some raw twitter data from Wang Yu, University of Rochester. With his generous help, the team plans to include this sentiment analysis on twitter data as a future work and improvement, since we didn't have enough time to fully apply.

Reference

1. *Pathways to news, Amy Mitchell, Jeffrey Gottfried, Michael Barthel, and Elisa Shearer, http://www.journalism.org/2016/07/07/pathways-to-news/*

2. *2016 the state of the first amendment survey, Newseum Institute's First Amendment Center, http://www.newseuminstitute.org/first-amendment-center/state-of-the-first-amendment/*

3. *Media Bias, Student News Daily, https://www.studentnewsdaily.com/types-of-media-bias/*

4. *Media bias in presidential elections: A meta-analysis, D'Alessio and M.Allen, Journal of communication,http://jonathanstray.com/papers/Media%20Bias%20in%20Presidential%20Elections.pdf*

5. *Ranking the media from liberal to conservative based on their audiences, The Washington Post, https://www.washingtonpost.com/news/the-fix/wp/2014/10/21/lets-rank-the-media-from-liberal-to-conservative-based-on-their-audiences/*

6. *Understanding Sentiment Analysis: What It Is & Why It's Used, Kristian Bannister, https://www.brandwatch.com/blog/understanding-sentiment-analysis/*

7. *What Donald Trump was up to while John McCain was a prisoner of war, Michael E. Miller and Fred Barbash, Washington post, https://www.washingtonpost.com/news/morning-mix/wp/2015/07/20/what-donald-trump-was-up-to-while-john-mccain-was-suffering-as-a-prisoner-of-war/?utm_term=.a89aa87008f6*

8. *Candidates' Speech Analysis, Erica Kim, http://firefly454.github.io/projects/speech/index2.html*