2016 Presidential Twitter News Coverage

Final Project Description

Michał Siedlaczek
ms8982@nyu.edu

Project page on Github: *https://github.com/NYU-CS6313-SPRING2016/Group-19*
Video: *https://vimeo.com/167335847*
Working demo: *http://nyu-cs6313-spring2016.github.io/Group-19/*

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

News sources give coverage to candidates for various reasons: ratings and income generation, journalistic integrity, private agendas, etc. For example, it was leaked that Fox News was giving favorable coverage to Marco Rubio in the 2016 Presidential election in an attempt to help him defeat Donald Trump. Also, less-popular candidates like Rand Paul claim their lack of popularity is primarily driven by the dearth of media coverage. This visualization is intended to identify and visualize relative differences in media coverage and its sentiment on Twitter (in form of short status updates, called "tweets") in order to identify biases and interesting outliers.

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

- Is there a leading sentiment along all networks/candidates?
- Which candidates are being helped/hurt (in terms of positive and negative tweets) the most by each network?
- Do any networks disproportionately (dis)favor a particular candidate? If so, whom?
- How does the support of a network for a candidate change over time? Are there any significant or sudden changes and when?
- What might have been the reasons for sudden changes (if exist) in sentiment of the tweets?

What is your data about? Where does it come from? What attributes are you going to use? What is their meaning? What are their attribute types (data abstraction)? Do you plan to generate derived attributes? If yes, which and why?

Note that the names of the derived attributes are in *italic*.

Attribute name	Attribute type	Meaning	Value range / categories
Time	Ordinal	When the tweets were posted.	02/15/16[1]-05/18/16
Network	Categorical	What network posted the tweets.	FoxNews, CNN, MSNBC, BBCWorld, ABC, CBSNews[2]
Candidate	Categorical	What candidate(s) the tweet was about.[3]	Trump, Cruz, Kasich, Clinton, Sanders, Rubio, Bush, Carson, Rand
Sentiment	Quantitative	Sentiment of the tweet.[4]	$s \in \{-2, -1, 0, 1, 2\}$
Retweets	Quantitative	The number of retweets of the tweet.	Integer $n \in (0, +\infty)$
Impact	Quantitative	The impact of the tweet = *sentiment* × *retweets*	Real number $x \in (-\infty, +\infty)$

What have others done to solve this or related problems?

http://www.r-bloggers.com/presidential-candidate-sentiment-analysis/

The author of the above article analysed the positivity and negativity of the the tweets posted by the candidates themselves.

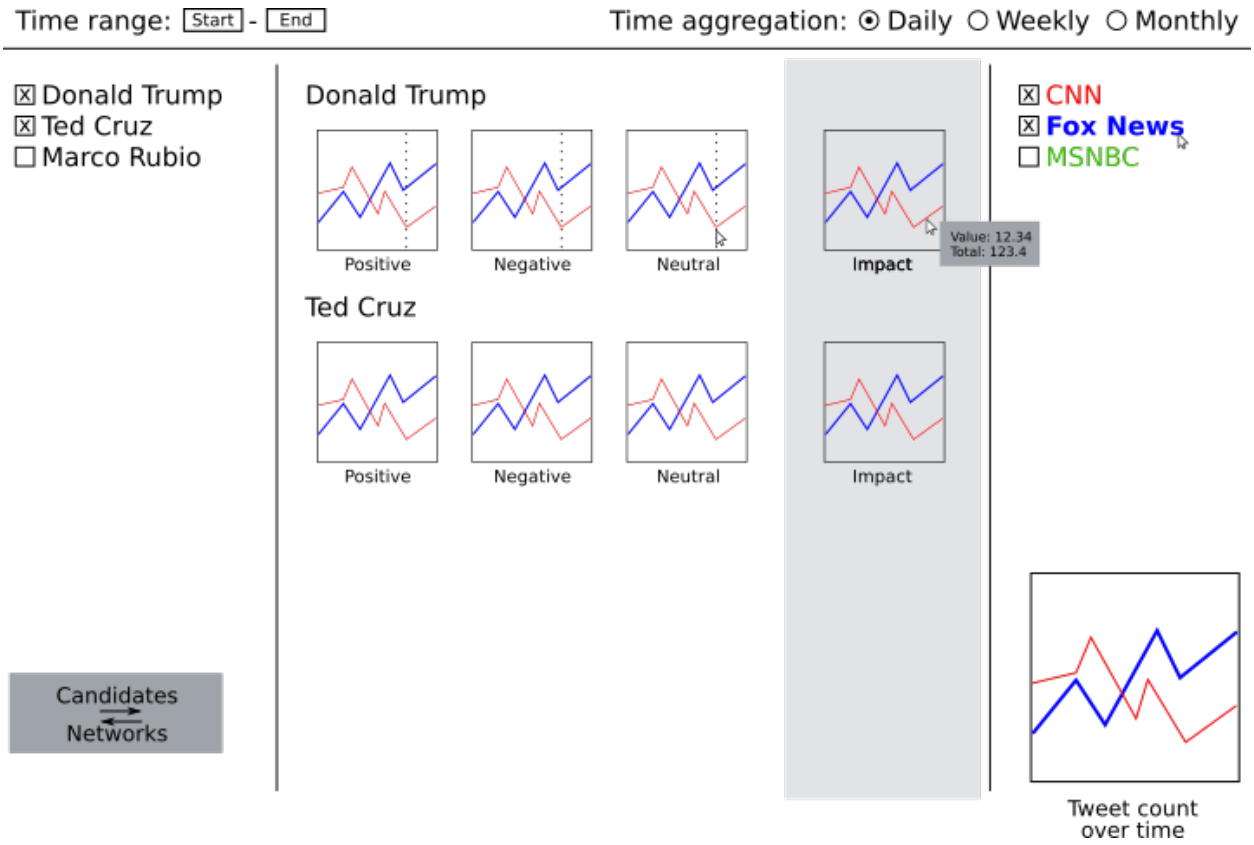
[1] With the exception of FoxNews: unfortunately, the limitations of Twitter API made it impossible to retrieve FoxNews' tweets from the first half of March

[2] These are the names of the Twitter channels.

[3] A simple filter is used to determine that information based on keywords and substrings. A tweet can be assign to multiple candidates.

[4] Stanford CoreNLP (`http://stanfordnlp.github.io/CoreNLP/`) is used to assign sentiment scores.

Initial Mockup



Description

The visualization is intended to show changes in media coverage of presidential candidates over time. Although the data is limited to some time period, it is possible to further zoom in by setting a different start and end date (the top left corner of the page).

The charts show quantitative data aggregated by specific period of time. It is possible to aggregate by: day, week, and month. The top right toolbar sets the parameter.

Below, the page is divided into three sections. The side columns represent two categorical dimensions. The first one is a list of all analyzed candidates, the other—a list of all media outlets whose tweets are analyzed. Both dimensions can be filtered by choosing one or more values we wish to analyze.

The middle section contains a list of rows, each of which represents one candidate. For each candidate, four charts are displayed. The first three of them show the change of number of tweets (positive, negative, neutral, respectively[5]) by media outlets in time. Each color-coded line represents one media outlet (also color-coded on the list). All of those share a common scale. The last chart shows impact over time. Impact on a given day (week/month) is a value derived from real sentiment scores multiplied by the number of retweets (a measure of significance or popularity of a tweet). This chart has a different scale from the others, that is why it is seperated graphically.

[5] A tweet is considered to be positive (negative) if it has any positive (negative) score. The tweets that are neither positive nor negative are considered to be neutral.

Additionally, the color-coded dimension (media outlets) is summarized by a simple chart showing tweet count for a given network over time. The scale of this chart does not have to match the scale of the charts in the middle section. However, it is important to note that each chart under a candidate's name needs to have a scale matching all the charts below. This is so comparison between candidates is possible.

Interaction

Hovering the coursor over an item in the networks list should highlight all color-matching lines in all the charts by making them thicker. Checking or unchecking a candidate causes displaying or hiding its row in the middle section. It does not affect the bottom right tweet count chart.

Checking or unchecking a network causes recalculation of scales and displaying the network's line along the other ones in all the charts.

Another interaction is hovering over one of the three volume charts (positive, negative and neutral tweet counts). A dotted line is drawn in the same place to make comparing values easier.

When hovering over a line on any chart, a pop-up is displayed with the value at this point, and total sum over the whole period (between start and end time).

The order of displaying the candidates can be changed by drag-and-drop on the left list.

The bottom left button flips the categorical dimensions. This means that the networks are on the left, and the candidates are on the right now. The rows in the middle correspond to the networks, and the lines to the candidates, which are now color-coded. Also, the bottom right chart shows the counts of tweets related to particular candidates.

Explanation

Coverage difference, as well as overall media coverage, of networks and candidates can be easily read from the bottom right chart.

The impact of media on candidates can be seen in the middle section, on both volume and impact charts. We can observe how support of different media outlets changes over time on volume charts. Impact chart shows estimates on how much it might affect general public. With some luck, we could link sudden changes in support to events happening at that time, like a particular primary or debate.

We can compare overall information about networks as well as candidates. Flipping those two dimensions helps analyzing the changes and differences in two different contexts.

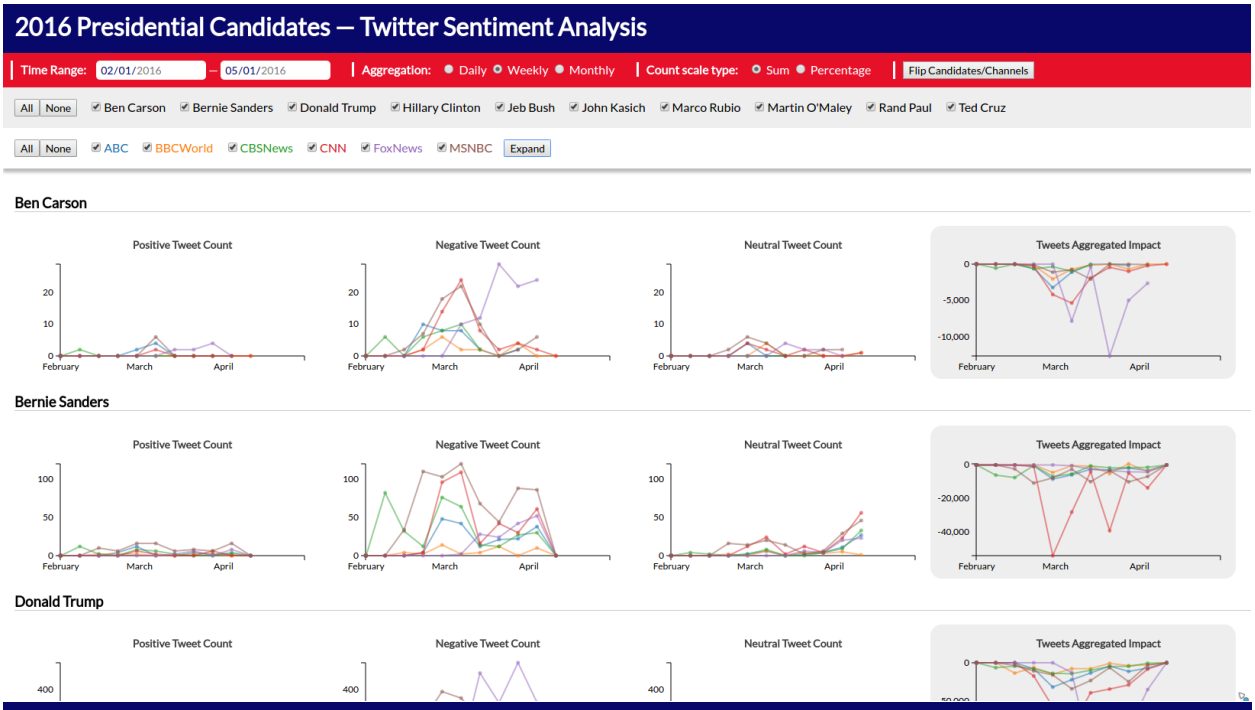
Manipulating the time aggregation type, can show some patterns that occur in time, and tell us how quickly they change. Zooming in by setting the time range allows us to look closely at particular interesting time periods, e.g., when important political events happen or vital news break out.

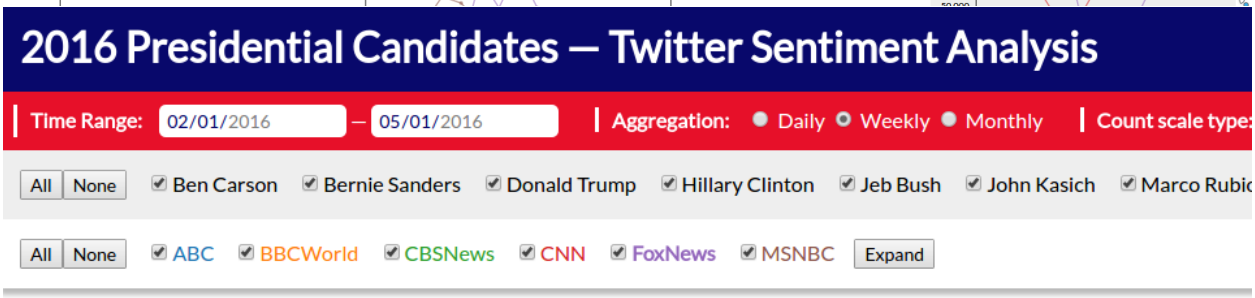
Project Update

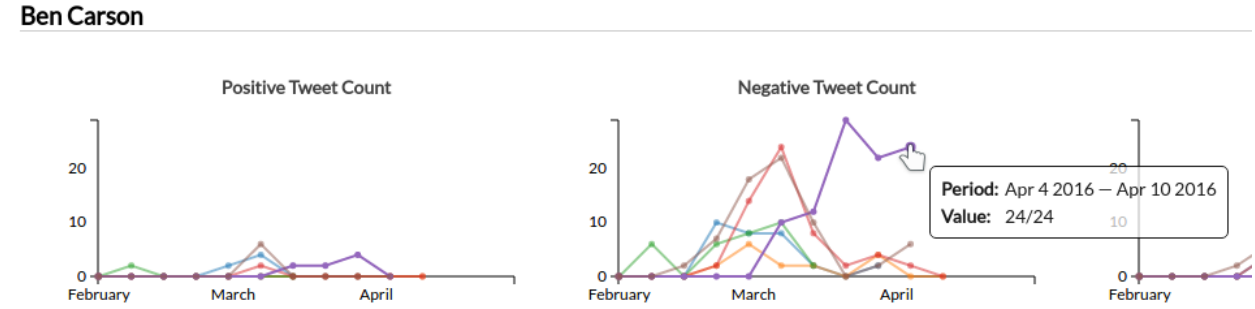
These are the major changes I decided to make:

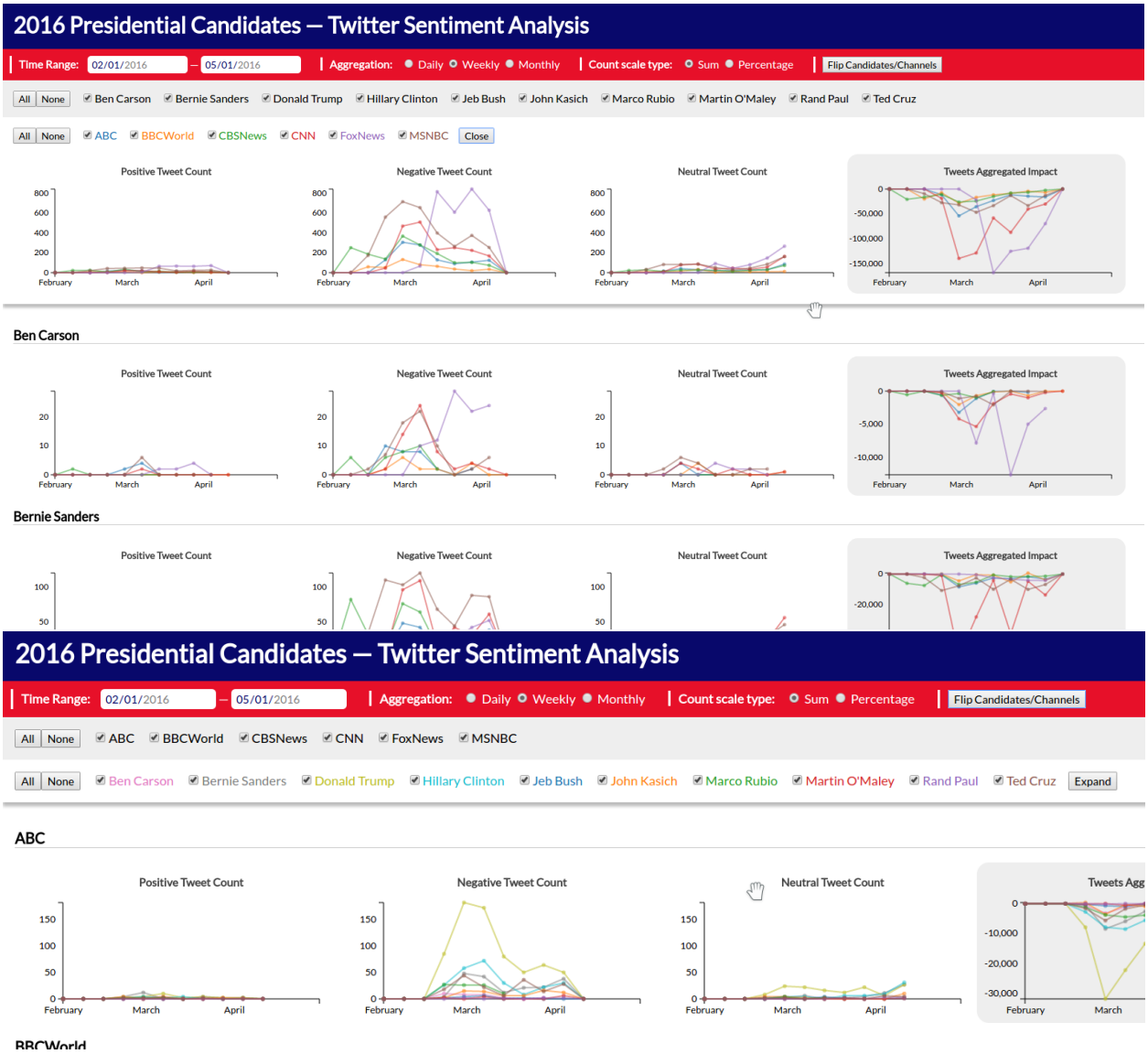
- I moved both dimentions (initially left and right menus) to the top and made them horizontal instead. The reason is to save up some room on the sides, and therefore increase the resolution of the entire visualization.

- For the same reason I moved "Flip Candidates/Channels" button to the toolbar.
- I added "Expand" button next to the second list. It shows global statistics (see the explanation and pictures below).
- Each row's Y axis is scaled separately. Instead of a common scale, I introduced an option to show the tweet volume in percentages (top menu).









The first picture shows an overview of the vizualization. At the top, there is a red toolbar containing all the settings: (1) **time range** for which the data is shown, (2) time periods by which the data is **aggregated**, (3) the **scale type** for tweet volume (it is possible to show either counts or portions of all the tweets for the time period), and (4) the **flip button** that switches *candidate* and *channel* dimensions places (the explanation below).

The next bar shows the list of all the candidates. They can be filtered out using the checkboxes. Each checked item corresponds to a row of charts below.

Next, there is the list of all the channels on Twitter. Each of the checked items corresponds to one line in each chart below. Both the names in the list and the lines are color-coded. Moreover, when you hover over either one, the name is displayed in bold, and the line's opacity is set to 1 (as opposed to 0.5 as default). You can observe this in the second screenshot. The other thing you can observe is a tooltip that is displayed when the mouse cursor hovers over a point. The tooltip shows the period for which the aggregation is done, and the value (count, percentage, or impact score).

Each row presents four charts. The first three show tweet volume: each point is the number of tweets (negative, positive and neutral, respectively) about the candidate posted by the channel in the time period.

The last chart shows impact scores over time.

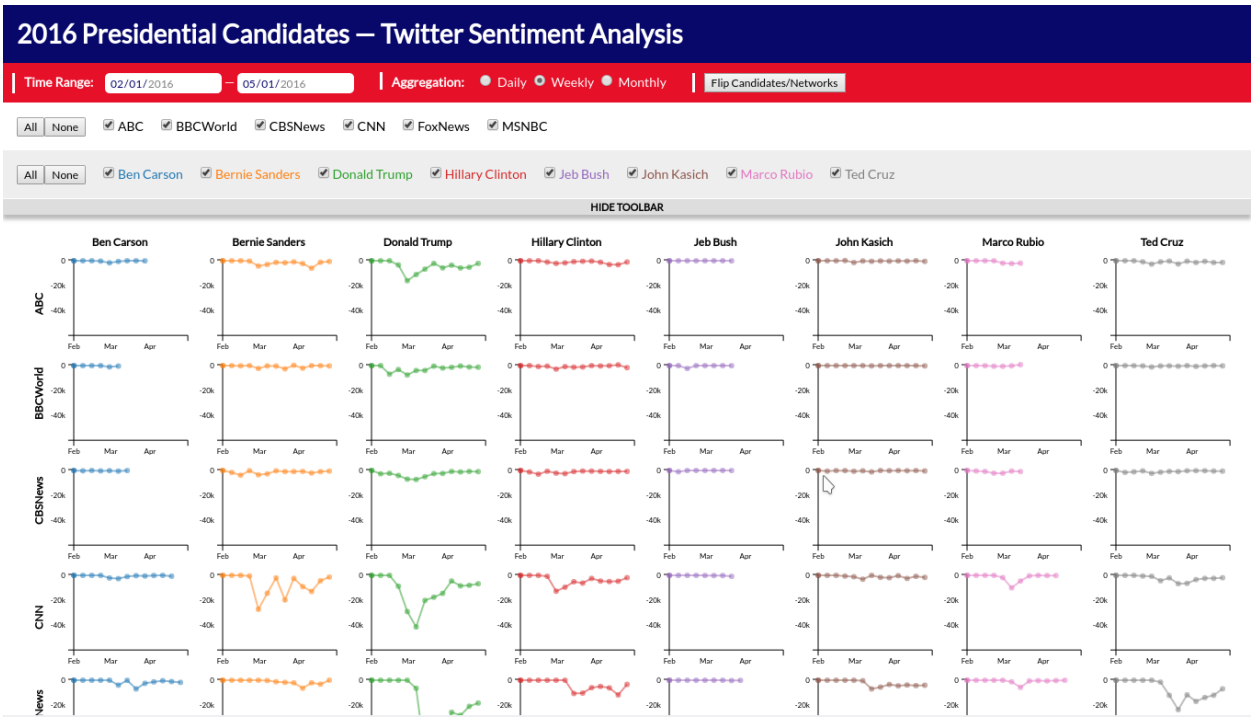
At the end of the list of the channels, there is “Expand” button. It expands a panel seen in the third picture. The set of charts is the same as for any row, except the values are cumulative for all candidates, thus they present global values over time.

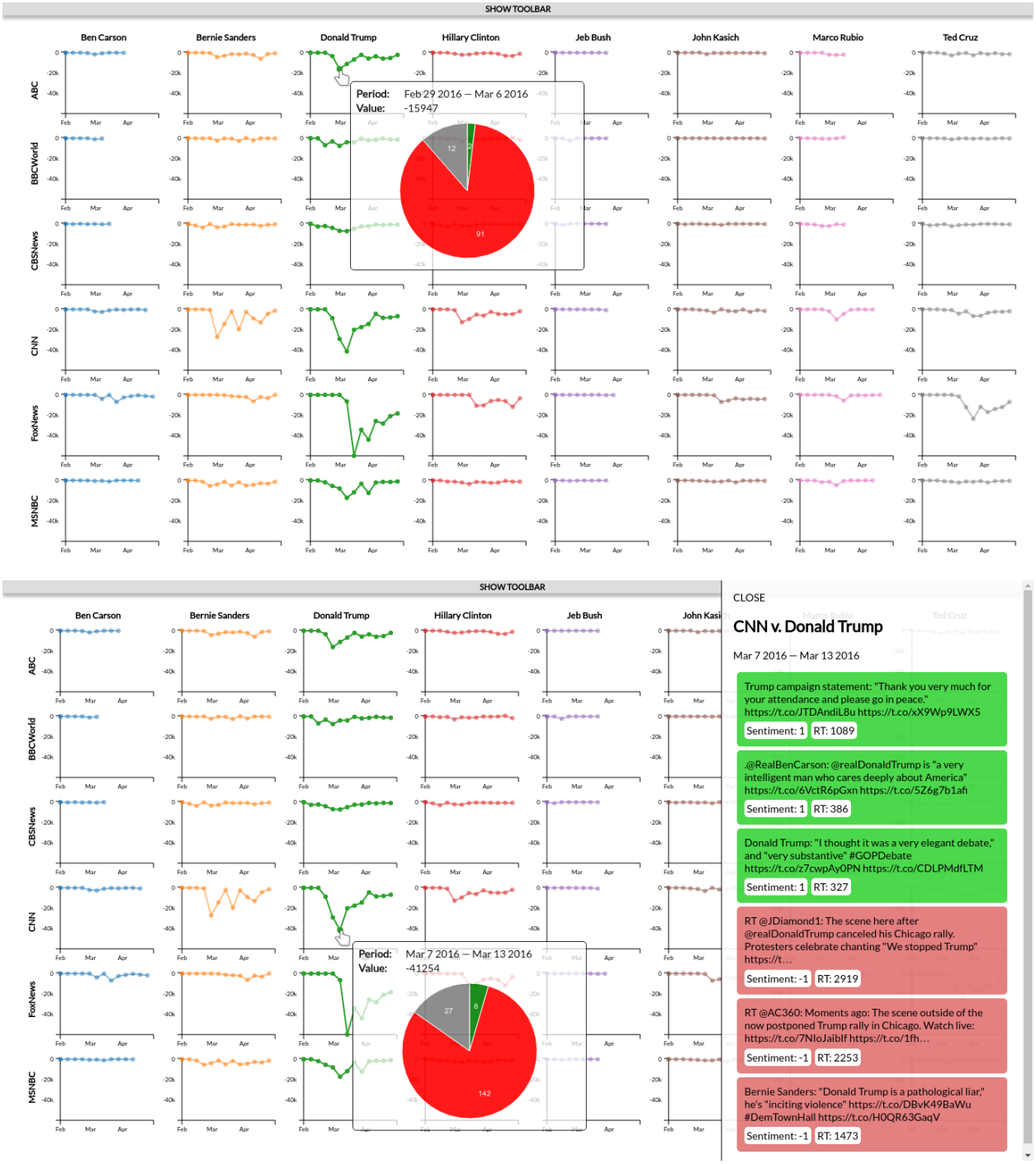
The last feature, the outcome of which can be observed in the last screenshot, is dimension flip. After clicking on the last button on the toolbar, the dimensions of candidates and channels simply switch places. It means that each row now corresponds to a channel (instead of a candidate), and, similarily, each color-coded line corresponds to a candidate now (as opposed to a channel). It impacts the global values (accessible after expanding the panel) as well.

Final Vizualization

Changes I made in the final vizualization:

- Information was difficult to catch at a glance, it involved to much scrolling:
  - I eliminated tweet counts and left only impact as the most interesting attribute.
  - I eliminated the global statistics (expandable panel).
  - I changed the layout so everything could be visible on the screen: a small multiples view.
- In the tooltip displayed when hovering over a point, I added a pie chart breakdown of the number of positive/negative/neutral tweets for that (network, candidate, period).
- I added a pop-up window that appears when a point on a line chart is clicked. It shows up to three most impactful negative and positive tweets for that (network, candidate, period).





The first picture shows an overview of the vizualization. At the top, there is a red toolbar containing all the settings: (1) **time range** for which the data is shown, (2) time periods by which the data is **aggregated**, and (3) the **flip button** that switches *candidate* and *channel* dimensions places.

The next bar shows the list of all the candidates. They can be filtered out using the checkboxes. Each checked item corresponds to a row of charts below.

Next, there is the list of all the networks on Twitter. Both the names in the list and the lines are color-coded. Moreover, when you hover over either one, the name is displayed in bold, and the line's opacity is set to

1 (as opposed to 0.5 as default). You can observe this in the second screenshot. The other thing you can observe is a tooltip that is displayed when the mouse cursor hovers over a point. The tooltip shows the period for which the aggregation is done, the impact score, and (if applicable) a quantitative breakdown of positive/negative/neutral tweets.

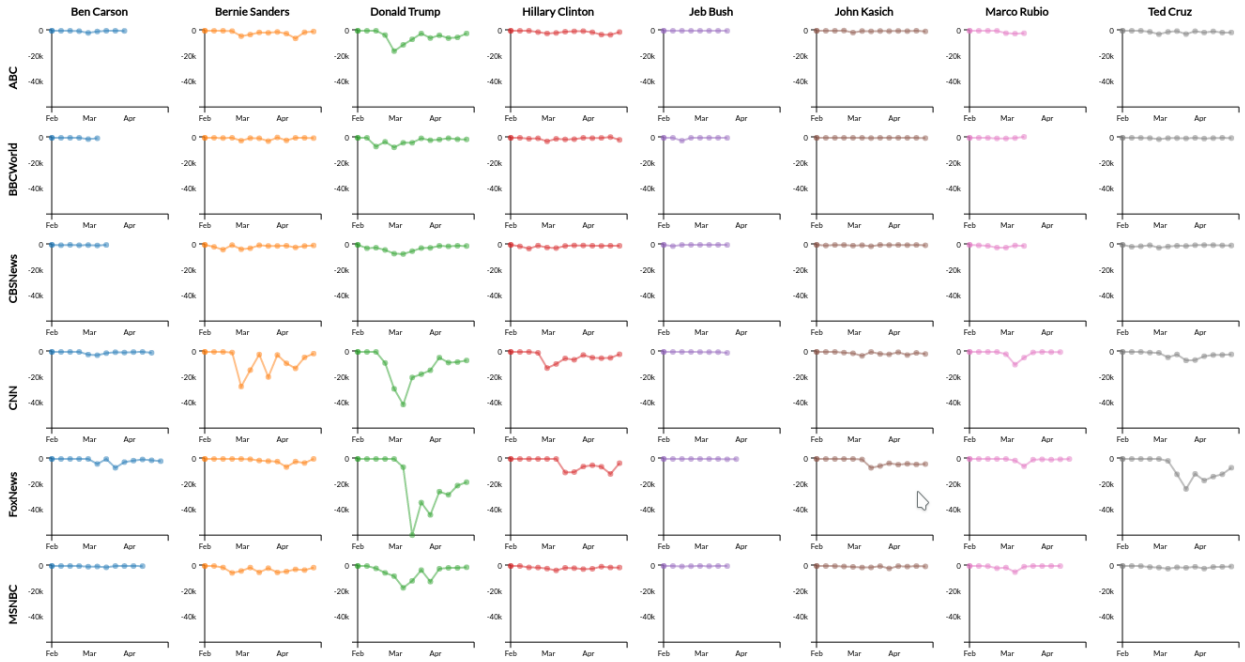
The main view is a grid of small multiples. Each row represents a network, and each column represents a candidate. Every line chart is a plot of changes of impact of tweets by a network in time (aggregated by a day/week/month). All charts have common scales.

Additionally, when a specific point on any chart is clicked, a pop-up window is displayed. It contains up to 3 most impactful (with the highest impact score) positive and negative tweets (along with their sentiment scores and number of retweets) for the chosen period.

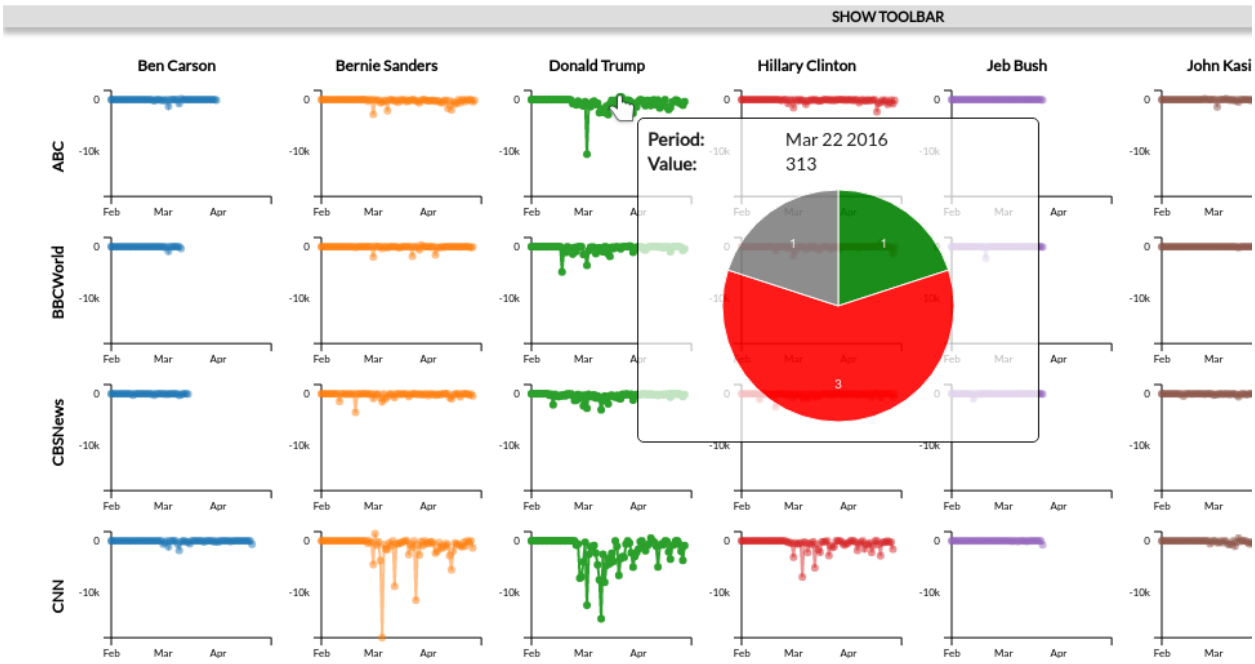
Data Analysis

This section describes data analysis. Let us look at the analytical questions and see what information we could extract.

Is there a leading sentiment along all networks/candidates?

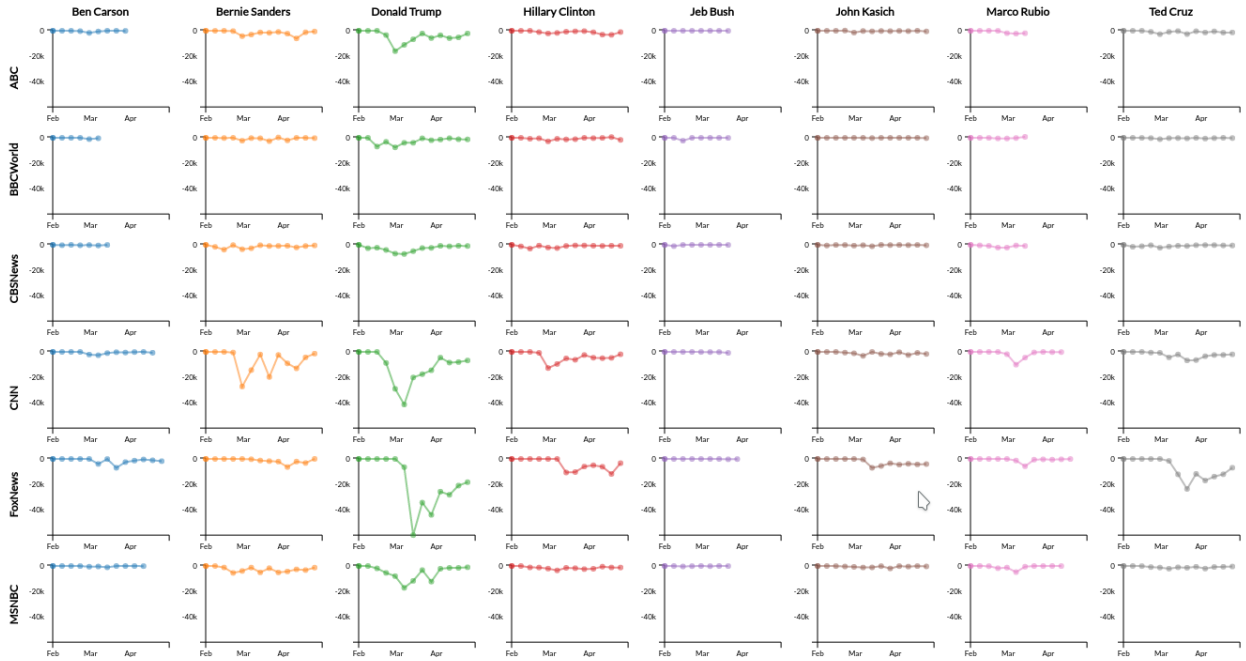


Just by glancing at the main view, it is immediately clear that negative sentiment dominates among all networks and candidates. In fact, there are very few data points with positive impact. To find those easier, we can change time aggregation to *daily*. However, even then there are very few days with impact above 0.



In the picture above, we can see that even though there was more negative than positive tweets, the impact is still positive. This is because of the higher number of retweets for the positive tweets that makes up for the higher impact score.

Which candidates are being helped/hurt (in terms of positive and negative tweets) the most by each network?



On the same view, we can also immediately see which candidates have been impacted the most. The most negative impact by far had the tweets about Donald Trump, especially those posted by CNN and Fox News.

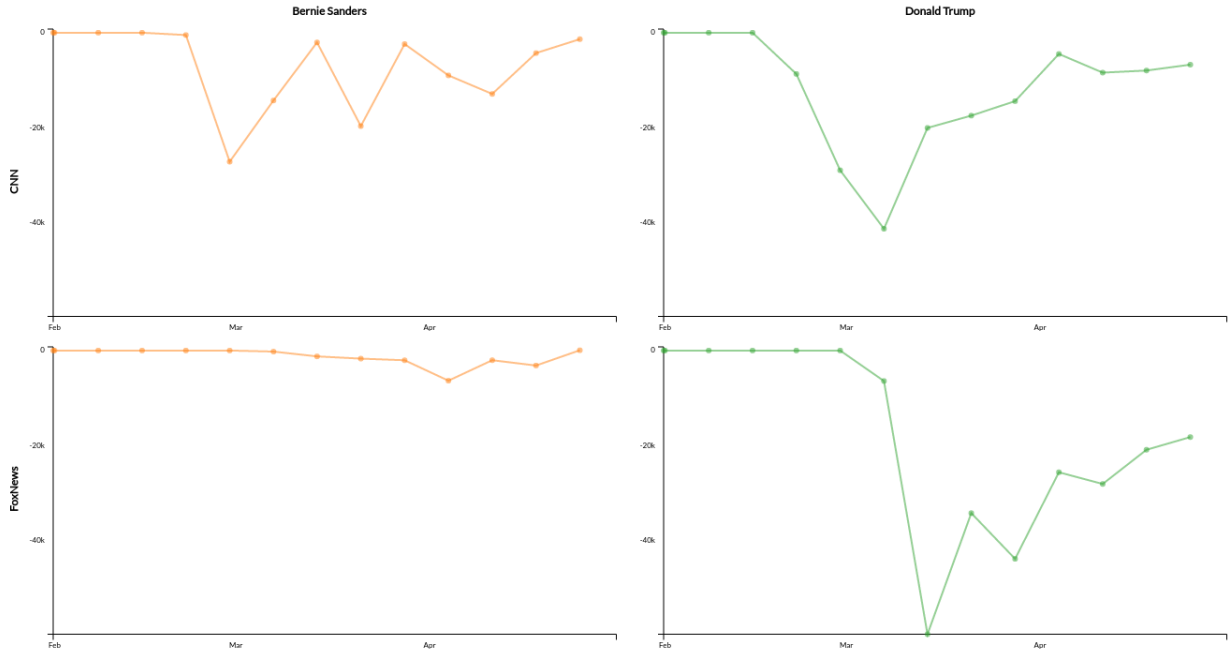
Following are Bernie Sanders, with negative impact mostly by CNN, Hillary Clinton, about whom most negatively tweeted CNN and Fox News, and Ted Cruz, with most negative impact from Fox News.

Do any networks disproportionately (dis)favor a particular candidate? If so, whom?

Going with our analysis further, we can see that BBC World and CBS News have the least impact, too little to tell whom they favor. Although, there is a slight dip in case of Donald Trump, his overall negative sentiment seems to justify it. A bit more apparent negative sentiment for Donald Trump is expressed by both ABC and MSNBC, however it still might very well be due to overall negative sentiment of the candidate. CNN clearly disfavor Bernie Sanders and Donald Trump; less apparent is the disfavor towards Hillary Clinton and Marco Rubio. Finally, Fox News clearly disfavor Donald Trump the most, following by Ted Cruz and Hillary Clinton, whereas Bernie Sanders was even able to get positive sentiment values.

How does the support of a network for a candidate change over time? Are there any significant or sudden changes and when?

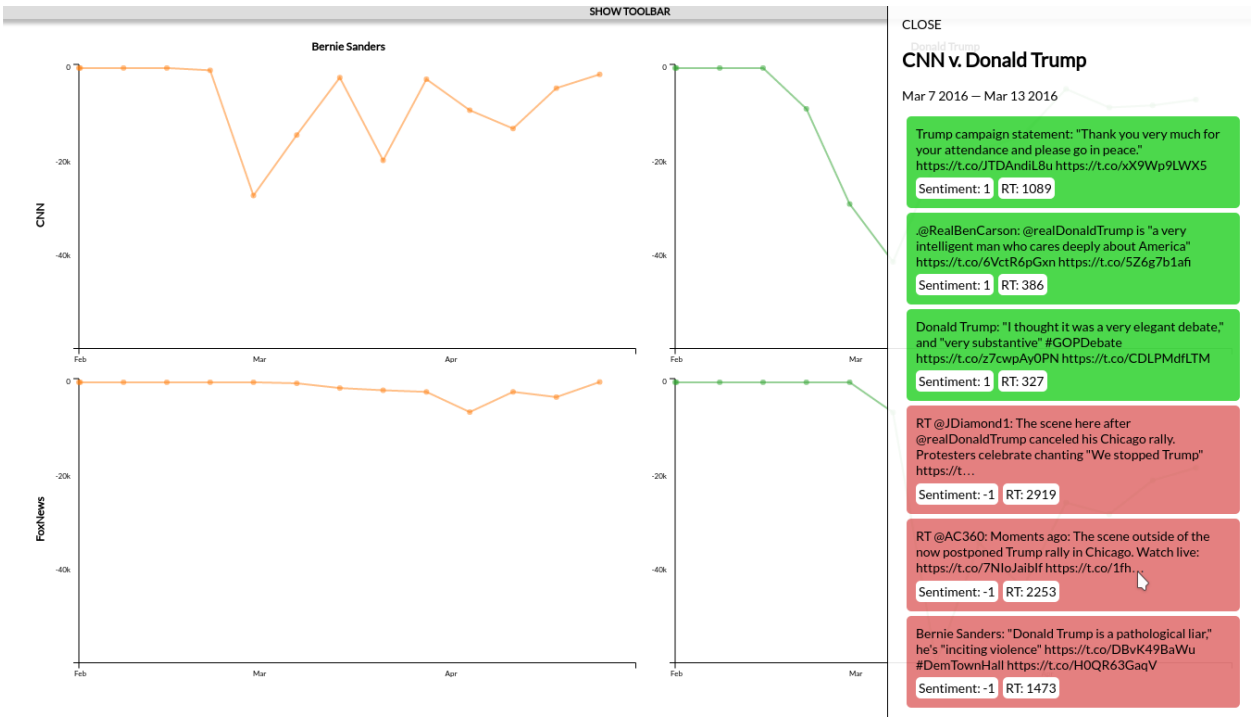
In the main view, we can see the most significant changes in case of Bernie Sanders and Donald Trump, especially when it comes to CNN and Fox News. To see it more clearly, let us limit the view to the above candidates and networks.



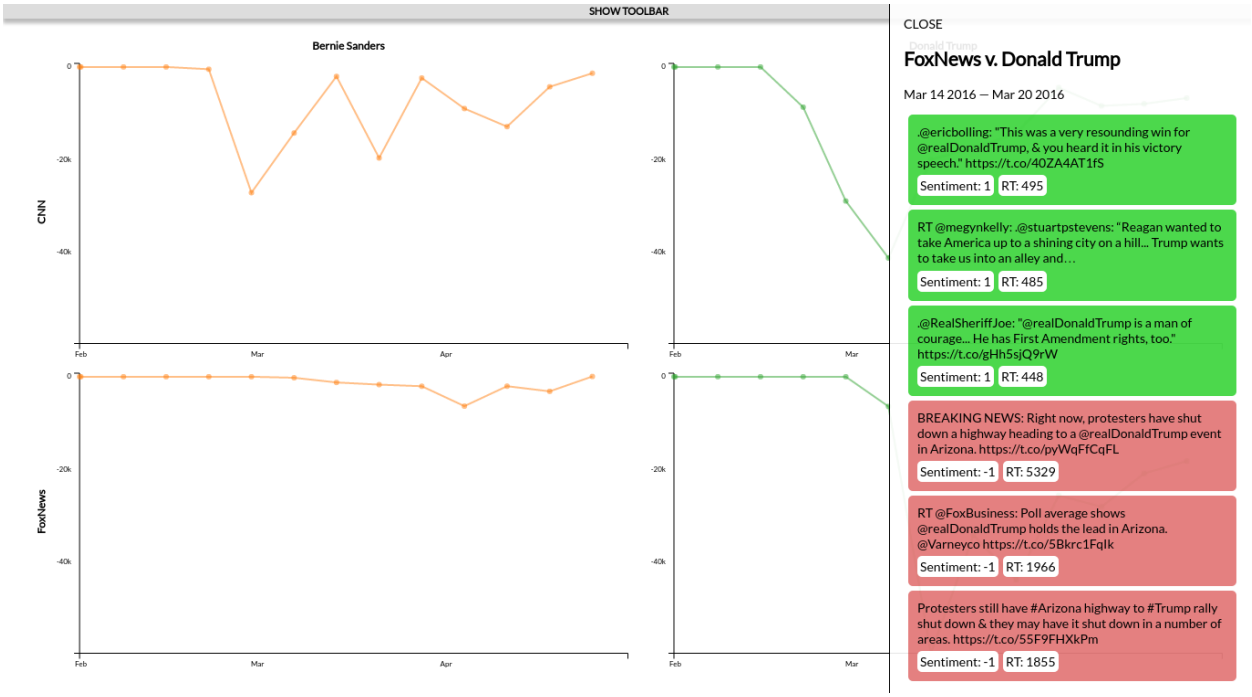
We can observe three major dips of sentiment of Bernie Sanders by CNN: early March, late March, and mid-April. However, the overall trend is increasing. In case of Donald Trump, there is one major dip in the first half of March, espressed by both CNN and Fox News. Later trend is also increasing.

What might have been the reasons for sudden changes (if exist) in sentiment of the tweets?

Having identified the trends and the point of interest, we can try and indentify the reasons for them. We can look up the most impactful tweets for the periods of question, focusing mainly on the negative ones, since we want to find out the reasons for negative sentiment. Although, it is difficult to identify anything specific about the weeks in which Bernie Sanders had the negative dips, we can tell some interesting stories about Donald Trump.



The tweets by CNN tell a story about Donald Trump cancelling his rally in Chicago and protesters trying to stop it (successfully). Whereas the most impactful tweets by Fox News mention a shutdown of a highway in Arizona due to people protesting against Trump's rally in the state. Those two event might have a big impact on the public, considering the number of the retweets and the overall sentiment for that period of time.



Limitations and Future Work

- It would be good to be able to analyze sinigle candidates (networks) one by one, just by clicking "next" to change to the following candidate (network).
- Also, a function to joing multiple columns together would be useful, in order to see multiple lines in the same chart for easier comparison.
- Now, all charts have a common impact scale. Furthermore, it is calculated globally and even when some charts are hidden, the scale is not recalculated. In order to see more subtle changes in some of the charts, it would be useful to allow to "zoom in" by recalculating the scales for individual plots or simply recalculating scales each time a network or a candidate is shown or hidden.