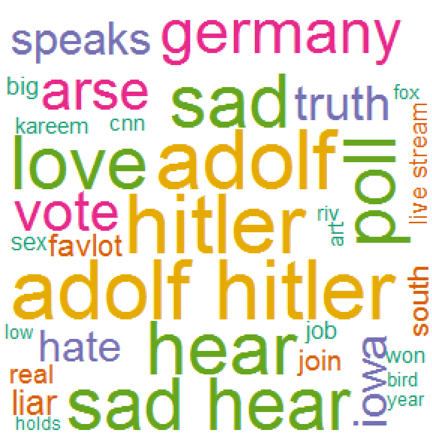
PHAN, Minh; FASIH, Mahir; DELA EVA, Regina

TWITTER TOPIC CLUSTERING

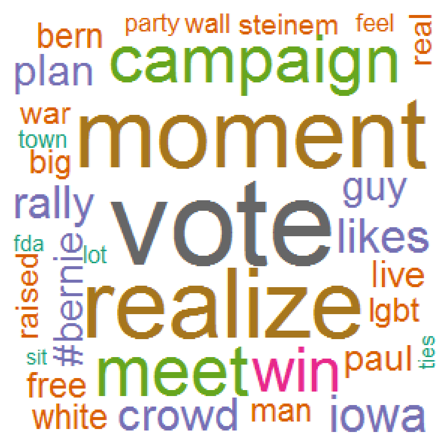
Tweets on Candidates for the U.S. Presidential Election 2016

DONALD TRUMP

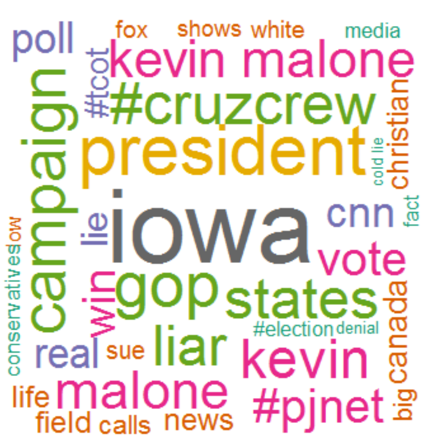


HILLARY CLINTON

BERNIE SANDERS



TED CRUZ



MARCO RUBIO

EXECUTIVE SUMMARY

The purpose of this exercise was to discover topics from tweets about each of the 5 selected candidates.

The desire to find meaning in constantly evolving political discussion using Twitter data was a very challenging project. The process and its challenges will be discussed in the next few sections. I was a very useful learning curve for us and we came to the conclusion that k-means clustering was quite limited to provide us with very meaningful topic clusters.

Due to the diverse and non-uniform nature of data and the unavailability of prior labels to test with, we believe that the topic clusters we created provide us some insight on the "word on the street" about each U.S. Presidential candidate, but we discovered that there is significant room for improvement. These improvements will be discussed in the last part of the report.

Despite not finding clusters that could generate political discussion, there were some interesting insights that we did come across:

- **Bernie Sanders** is the only candidate with issue oriented clusters (taxes, health etc.).
- The word **Liar** appears many times in tweets associated with **Ted Cruz**.
- A cluster from the Test data set likens **Donald Trump** to a Facist historical figure.
- It's not smooth sailing for **Hillary Clinton**, there's a lot of mention of her email scandal and Bengazi hearings.
- Overall, all candidates had significant tweets on Polls, Endorsements, Debates and Primaries/Caucuses.

APPLICATION FUNCTIONALITY

The project was split up in four phases and an R-script is submitted for each phase:

1) **Module 1: Download the data**
   Initially, 100,000 tweets were downloaded from twitter using the twitteR package on 27th January 2016, and from there 10,000 tweets were downloaded daily and saved as CSVs for each of the 5 candidates separately until 17th February 2016. Finally, we have created a database of 1,500,000 tweets for 5 candidates.
2) **Module 2 & 3: Merge & Clean the data**
   During this process only relevant fields were kept, duplicate tweets were removed and tweets were sorted by date. The next step was to remove non-ASCII characters, hyperlinks, twitter jargon (via, RT), extra spaces, and converting all text to lowercase in data.table. The rest of the clean-up process was conducted in a corpus where we removed numbers, punctuation (preserving hash-tags), and removing a standard and custom list of stopwords. At each stage of the cleaning process a new column of clean text was created.
3) **Module 4: Topic clustering**
   During the clustering step, the tweets were split into training and test data based on the day of the week on which they were published. Mon, Tue, Thu, Fri and Sat for ***training data set***, and Wed & Sun for ***testing data set***. The document term matrix was created with 1-grams and bi-grams and a TF-IDF weighting. After conducting a frequency and sparsity analysis, Sparsity was set to 0.99, and that reduced the dimensions from over 100,000 to around 90.
   In order to implement k-means, the number of clusters were determined using the elbow method. Based on the k, the k-means function was applied to create k clusters and the top 5 centers (terms) were set to be the "topics" for each cluster. This process was repeated for the test data set to compare the results, and then for each candidate.

The process described above was run separately for each candidate to determine their topic clusters. At no point was data from one candidate merged with the data from another candidate. In the Appendix the process map that outlines the process described above can be found.

CHALLENGES MET DURING THE PROCESS

#	Challenge	Solution
1	**Install package, hidden/missing package?** There were some packages like RWEKA (used for n-grams), that did not available on CRAN and they were dependent on some other packages, like java, that had some complications with installation.	Research on StackOverflow.com and other forums provided with suitable steps to overcome this issue.
2	**Cleaning up tweets** An important step in trying to find topic information on political topics in tweets was to remove content that was considered to be noise and that did not have any meaning. In this case, it would have to do with hyperlink strings and text associated with retweets ("via", "RT", username of retweet).	Removing some of these items was not a very simple task. Instead of using standard corpus cleaning techniques, we resorted to Perl regular expressions.
3	**Processing times** Data-processing using the gsub function in a data frame data structure proved to be more challenging as the size of the data set increased (e.g. around 4 hours to clean-up). The processing time for 1.5 million tweets was enormous at the end of the project. Determining the number of clusters per candidate took up at least 15 minutes to process for each candidate. Naturally, the processing time increased when the size of the training data doubled (when we switched the scope of the training data from the first 100,000 tweets to tweets from 5 days a week).	To overcome this issue, we converted the data frame to a data.table before converting it into a corpus. To overcome the second issue, we realized the importance of the prototype of 5000 tweets. That was an incredibly fast and reliable way to test our code before running it on 1.5 million tweets and wait hours for the process to end.
4	**Removing Unicode characters** We initially tried to use the iconv() function in R to remove non-ASCII characters. But that function was resulting in other complications and not really cleaning up the data either.	In the end, we used Perl regular expressions to match on strings starting with "<U" and ending with ">" to target and remove those characters.
5	**N-Grams** We decided to include bi-grams as some bi-grams ("white house", "South Carolina") were occurring many times and had significant meaning when together. The challenge with that decision was to find a package that performed that task and also that we would be losing out on a lot of 1-grams if we included bi-grams.	During our research we discovered the RWEKA package that created n-grams. We decided to include both 1-grams and bi-grams. In the case where bi-grams were more significant than their 1-gram components the TF-IDF weighting and sparsity application would filter them out. The same would be applicable to all bi-grams that had no meaning.
6	**How many clusters should there be?** To determine how many topic clusters there should be, was the next logical challenge for us. When we applied the method to discover the number of clusters, it was not always very clear. Sometimes there were no elbows, and sometimes there were multiple elbows (See Appendix).	Thanks to the guidance provided by the professor, used the "elbow" method to identify the number of clusters k in our k-means topic clustering project. When the results from the elbow were not clear or in the case of multiple elbows, we had to test multiple Ks till we could find an optimal k.
7	**What does each cluster represent?** Near the end of the process we decided to exclude the names of all the candidates to give more meaning to our topics. Including names of other candidates meant that all the clusters would be skewed by the presence of the names, and no other	We chose top 5 centers (terms) from each of clusters to represent the topic of that cluster. This included 1-grams and bi-grams and provided some insight into what that cluster was about. (See Appendix).

	opinion. What we were looking for are not comparisons, but the "word on the street" on each candidate. That is why we removed the names. But even after that what did each cluster represent?	
8	**Evolution of clusters** After running a test of the whole process, we came to realize one major flaw in our process. The 100,000 tweets downloaded on January 27th 2016 as our training data set, did not seem to in line with the latest discussion. The problem here was that the test data set, downloaded a few days or weeks after the training data set would always be out of sync due to the nature of constantly evolving Twitter data and political developments.	We segmented the data by days of the week. Tweets from 5 days of the week served as part of the training data set (Mon, Tue, Thu, Fri, Sat), while tweets from the remaining two days (Wed, Sun) were part of the test data set; representing a mixture between the old and new topics, and providing a constantly evolving training and test data set to ensure that our topic clusters also evolve every week.
9	**Testing the model** The biggest challenge was to ascertain how to test the model. Since we did not have prior labels for the tweets in our test data set, we could not perform External Validation.	We wanted to use the test data set in some way, so we decided to repeat the process on the test data and compare the number and nature of clusters with those of the training data set (See Appendix).
10	**Nature of the data** Twitter data is very limiting when it comes to topic clustering. This is because each document is limited to only 140 characters. This results in a very low TF-IDF, and provides very little flexibility and insight when performing Sparsity and Frequency analysis.	A creative way to work with topic clustering and topic classification is discussed in the next section.

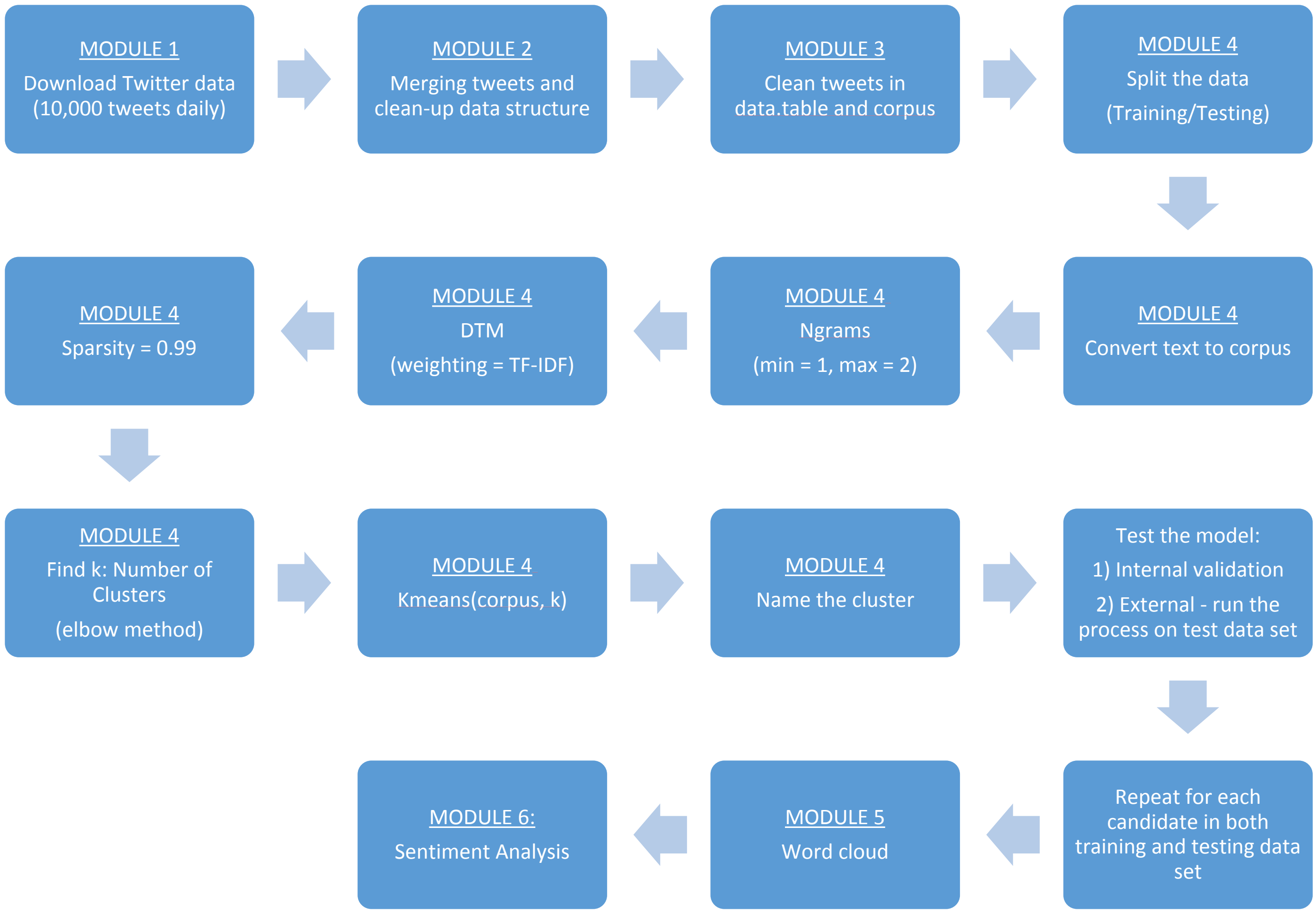
DESIDERATA FOR FUTURE WORK

A process has been created where the topic clusters evolve over time. This is our best take-away from the project at this stage. However, during the project we did discover other techniques and methods that would be very useful in making this project more accurate and meaningful. They fell out of scope for us, but they would be very useful to implement for other projects related to twitter data.

1) **Use of hash-tags (#) for external-validation and topic clusters**
   One of the main challenges we faced when trying to discover the theme for each candidate was that there were no standard topics, labels or "tags" that each tweet could be related to. Which is why we could not perform external validation, or topic classification and had to resort to clustering. One way to overcome the problem is to treat hash-tags as the labels and/or topics.
2) **Find info in hyperlink text**
   Another challenge was the limitation of 140 characters, and this makes the tweets not provide much information. Removing hyperlinks means removing hyperlink text. Sometimes there is useful information in the hyperlink text that can be salvaged to get more insight.
3) **Web crawling**
   Taking the case from (2) one step forward would be to actually go to the hyperlink pasted and use the content from there and add it to the tweet document. This would increase size of the document and bring more diversity to the document term matrices.
4) **KNN & Sentiment Analysis**
   Another project that we had started working on during the course of the topic clustering project was that of Topic Classification using KNN, and the possibility of performing a sentiment analysis as well. This would be possible if the tweets for each candidate were tagged by their name and then pooled together into one data set, to predict which candidate a tweet was referring to.

APPENDIX

PROCESS MAP



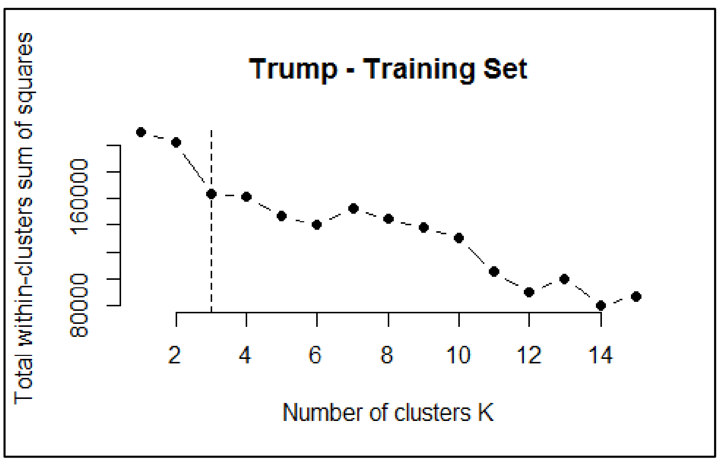
MODULE 1
Download Twitter data (10,000 tweets daily)
MODULE 2
Merging tweets and clean-up data structure
MODULE 3
Clean tweets in data.table and corpus
MODULE 4
Split the data
(Training/Testing)
MODULE 4
Convert text to corpus
MODULE 4
Ngrams
(min = 1, max = 2)
MODULE 4
DTM
(weighting = TF-IDF)
MODULE 4
Sparsity = 0.99
MODULE 4
Find k: Number of Clusters
(elbow method)
MODULE 4
Kmeans(corpus, k)
MODULE 4
Name the cluster
Test the model:
1) Internal validation
2) External - run the process on test data set
Repeat for each candidate in both training and testing data set
MODULE 5
Word cloud
MODULE 6:
Sentiment Analysis

RESULTS – TOPIC CLUSTERING

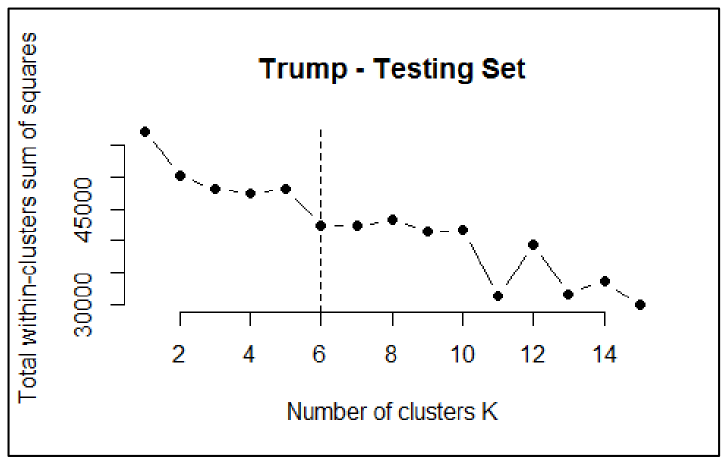
The next few sections are split up by candidates and their results; topic clusters and document distribution for both training and test data sets.

DONALD TRUMP (R)				
Training set			Testing set	
Cluster	Topic	Doc %	Topic	Doc %
1	love phoraone bruh kidnapped open letter	3.05	vote therapy guy guy therapy night	2.06
2	vote hampshire debate fox news	4.81	win time scalia vote republican	2.26
3	president president rt conchobar rt conchobar # realdonaldtrump	92.14	guy #realdonaldtrump #realdonaldtrump save #trump ’ sad	0.01
4			people support vote man hate	0.97
5			total liar total liar sunday presidential	1.07
6			president hitler adolf hitler adolf #realdonaldtrump	93.63

Trump - Elbow analysis on Training set

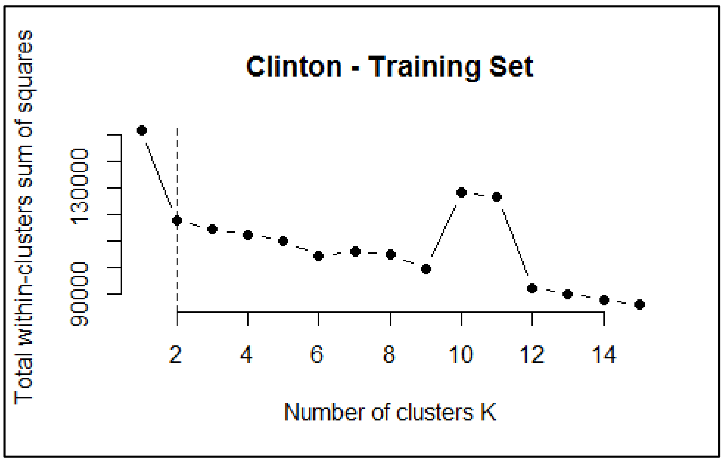


Trump - Elbow analysis on Testing set

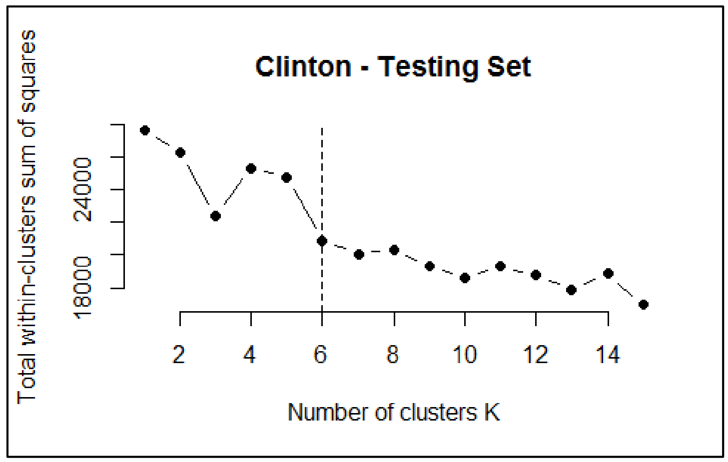


HILLARY CLINTON (D)				
	Training set		Testing set	
Cluster	Topic	Doc %	Topic	Doc %
1	candidate cnn women berniesanders #imwithher	3.05	donors state read nevada big	1.07
2	youth message latest latest message message youth	99.83	iowa win caucus hold voters	3.73
3			issue great deal prose great deal	91.28
4			deserve deserve black doesn’ deserve doesn’ black vote	1.43
5			town town hall hall dems hold	1.39
6			fbi email video emails benghazi	1.09

Clinton - Elbow analysis on Training set

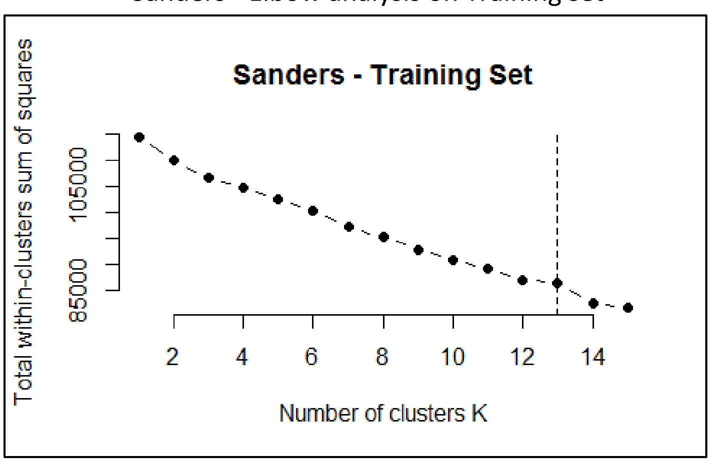


Clinton - Elbow analysis on Testing set

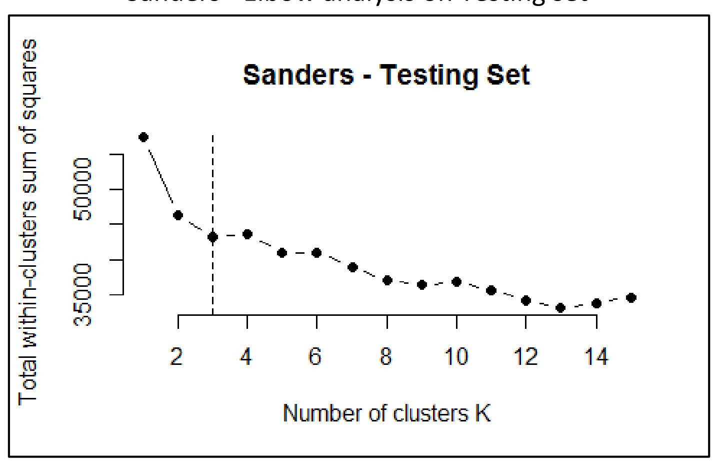


BERNIE SANDERS (D)				
Train			Test	
Cluster	Topic	Doc %	Topic	Doc %
1	campaign women health presidential people	2.0	hampshire win socialist million united	1.48
2	president obama love people supporters	76.17	million million january january raised million raised	1.68
3	bern iowa endorsement berniesanders win	1.02	vote realize moment moment realize realize vote	96.84
4	feelthebern hillaryclinton #bernie #demtownhall #berniesanders	4.07		
5	voting poll big democratic people	0.95		
6	vote young president #feelthebern make	1.28		
7	iowa caucus vote win rally	4.42		
8	berniesanders endorsement campaign hillaryclinton live	4.22		
9	taxes raise taxes raise video health	1.84		
10	win america iowa hampshire #feelthebern	0.50		
11	town hall hall town cnn democratic	1.54		
12	video campaign make president people	1.50		
13	support young people poll iowa	0.47		

Sanders - Elbow analysis on Training set

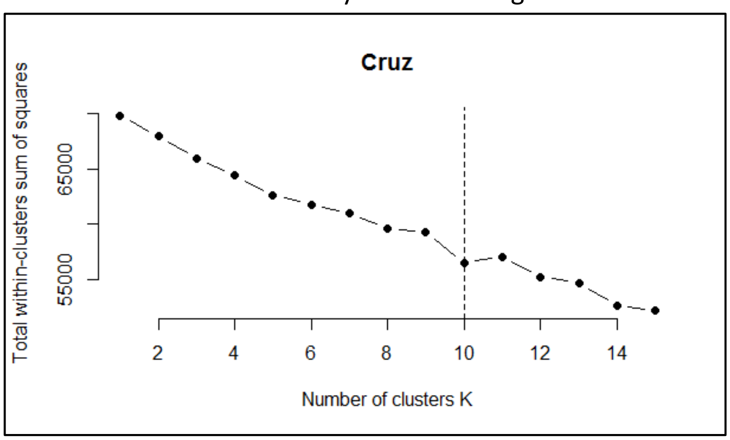


Sanders - Elbow analysis on Testing set

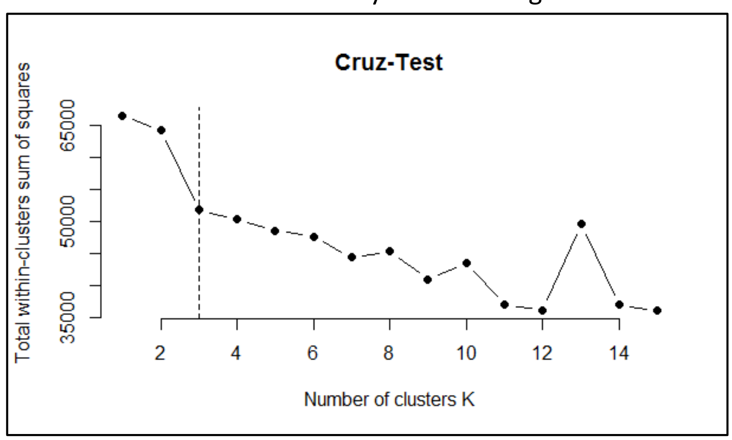


TED CRUZ (R)				
	Training set		Testing set	
Cluster	Topic	Doc %	Topic	Doc %
1	#cruzcrew #pjnet #tcot vote #scprimary	4.83	nasty dirty ballothold realbencarson ballothold tricks realbencarson ballothold lib… lib… realbencarson dirty tricks	1.15
2	presidential candidate conservative voters republican	2.35	debate debates beat debate … beat debates debates debate poll iowa	97.16
3	#trump #iacaucus realdonaldtrump united #cruz	1.18	tony perkins tony perkins endorses endorsement president endorsed leader	1.69
4	president endorses obama senator realdonaldtrump	1.07		
5	america #pjnet god christian #iowacaucus	0.88		
6	people #iowacaucus vote iowa obama	0.34		
7	poll cnn iowa liar trump	0.25		
8	cnn lies realbencarson lying liar	1.15		
9	iowa gop campaign liar states	86.83		
10	hampshire iowa #iowacaucus win debate	1.13		

Cruz - Elbow analysis on Training set

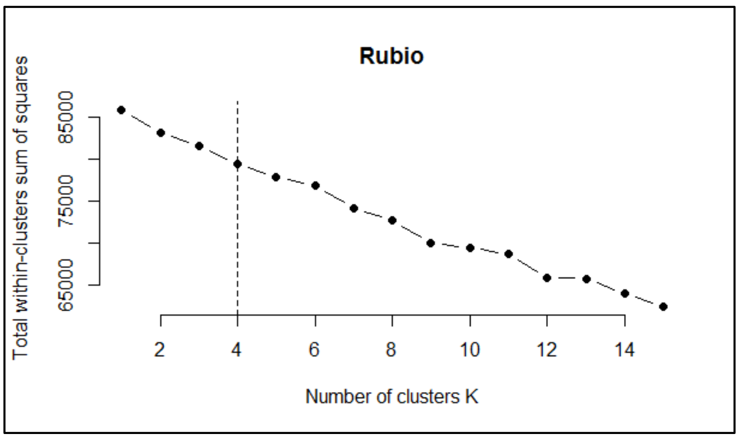


Cruz - Elbow analysis on Testing set

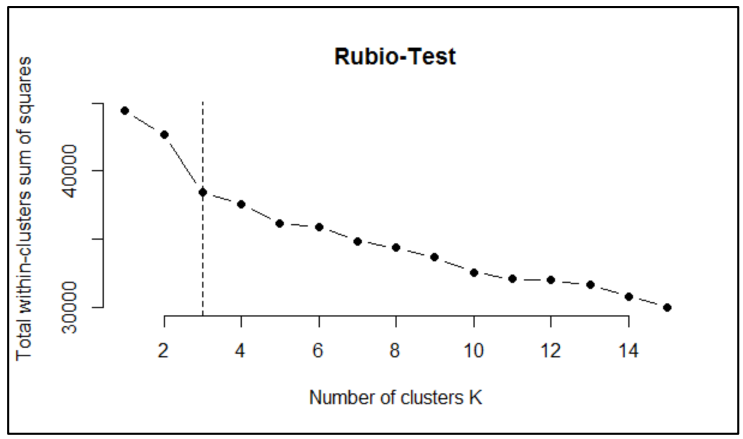


MARCO RUBIO (R)				
Training set			Testing set	
Cluster	Topic	Doc %	Topic	Doc %
1	iowa debate president campaign Obama	86.3	haley nikki haley nikki carolina south carolina south endorse haley endorse	4.68
2	moines des moines des register moines register	3.02	#gopdebate night people today realdonaldtrump immigration #rubio #teammarco	2.22
3	haley nikki haley nikki carolina south	4.91	debate iowa gop president register campaign hampshire obama	93.08
4	gop hampshire debate register endorsement	5.78		

Rubio - Elbow analysis on Training set

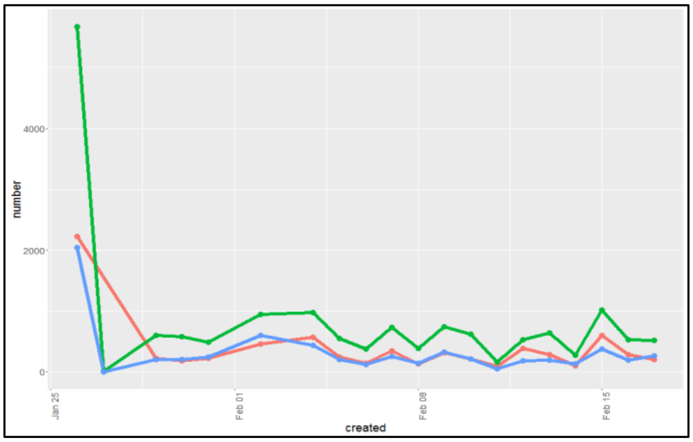


Rubio - Elbow analysis on Testing set

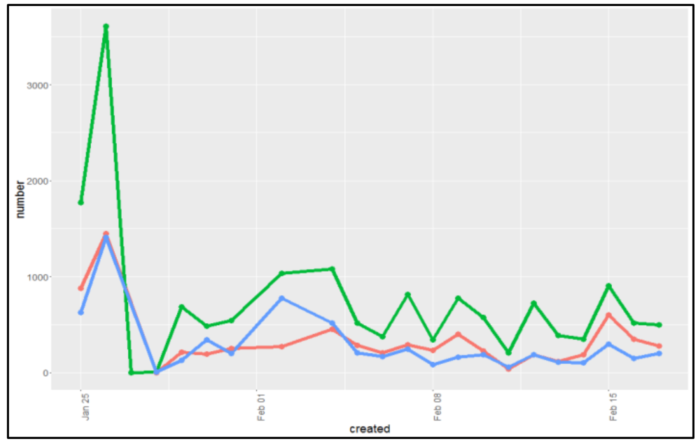


This result stimulated the sentiment of people who posted their tweets about these 5 candidates during the project period.

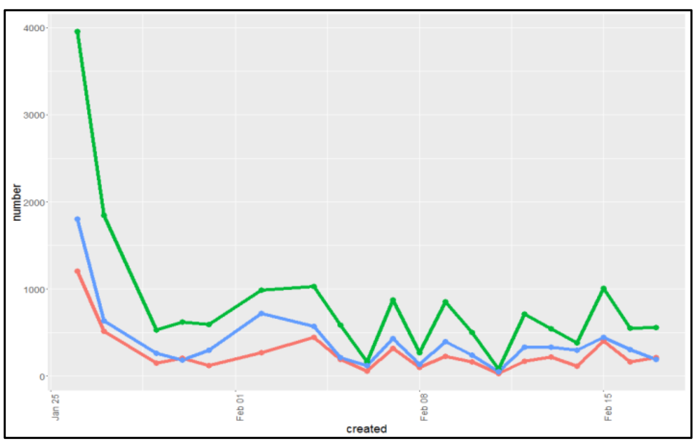
How did people feel about Donald Trump?



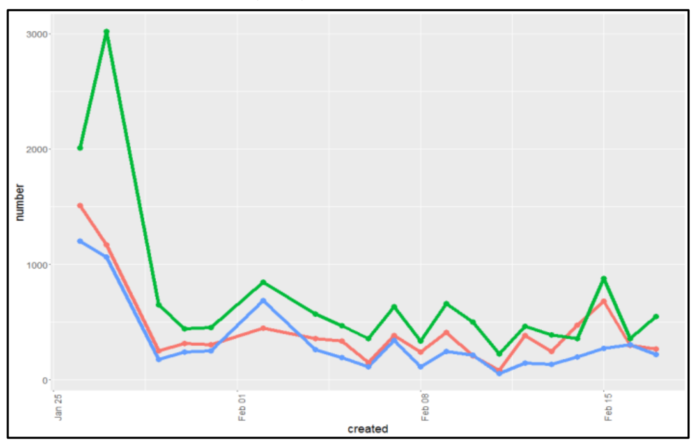
How did people feel about Hillary Clinton?



How did people feel about Bernie Sanders?



How did people feel about Ted Cruz?



How did people feel about Marco Rubio?