

12-1-2018

Using SAS™ software to enhance pedagogy for Text Mining and Sentiment Analysis using social media (Twitter™) data

Ramesh Subramanian

Quinnipiac University, ramesh.subramanian@quinnipiac.edu

Danielle Cote

Quinnipiac University

Follow this and additional works at: <https://scholarworks.lib.csusb.edu/jitim>



Part of the [Business Intelligence Commons](#), [Curriculum and Instruction Commons](#), [Management Information Systems Commons](#), [Online and Distance Education Commons](#), [Social Media Commons](#), and the [Technology and Innovation Commons](#)

Recommended Citation

Subramanian, Ramesh and Cote, Danielle (2018) "Using SAS™ software to enhance pedagogy for Text Mining and Sentiment Analysis using social media (Twitter™) data," *Journal of International Technology and Information Management*. Vol. 27: Iss. 2, Article 4.

DOI: <https://doi.org/10.58729/1941-6679.1380>

Available at: <https://scholarworks.lib.csusb.edu/jitim/vol27/iss2/4>

This Article is brought to you for free and open access by CSUSB ScholarWorks. It has been accepted for inclusion in *Journal of International Technology and Information Management* by an authorized editor of CSUSB ScholarWorks. For more information, please contact scholarworks@csusb.edu.

Using SASTM Software to Enhance Pedagogy for Text Mining and Sentiment Analysis Using Social Media (TwitterTM) Data

Published as Pedagogical/Teaching Case

Ramesh Subramanian

ramesh.subramanian@quinnipiac.edu

Danielle Cote

danielle.cote@quinnipiac.edu

Quinnipiac University

Hamden, CT, USA

ABSTRACT

This pedagogical paper describes how a graduate course in Text Mining was developed and taught in a fully online format at Quinnipiac University. The software used was SASTM Enterprise Miner. This paper discusses the design, software used and the methodology followed in the course. A critical component of the course required the students to delve deep into social media data by completing a detailed project on analyzing sentiment analysis using large files of social media data. A sample report of this project, which was a key deliverable for the course, is described at length in this paper.

KEYWORDS: Data Mining, Pedagogy, Sentiment Analysis, SAS, Course Design

PEDAGOGICAL STUDY INTRODUCTION

This pedagogical paper describes how a graduate course in Text Mining was developed and taught in a fully online format at Quinnipiac University. The software platform used was SASTM Enterprise Miner. This paper discusses the design, software used and the methodology Followed in the course. A critical component of the course required the students to delve deep into social media data by completing a detailed project on analyzing sentiment analysis using large files of social media data. A sample report of this project, which was a key deliverable for the course, was prepared by two of the co-authors, and is described at length in this paper.

BACKGROUND

Quinnipiac University, located in Hamden, CT, started offering a fully online MS program in Business Analytics (MSBA) in 2014. A total of 36 credits is required to complete the program, of which 24 credits come from core courses and 12 credits come from electives. All courses are of 7-week duration, and are offered through the university's Blackboard Course Management System (CMS). During the design phase of the program, it was decided that our program would be based on the SAS platform. This was done for both practical as well as strategic reasons. It was practical, because the SAS analytics software platform had good name-recognition in the industry. Moreover, there was a recognition that most of the CIS Dept. (where the MSBA was located) faculty were new to business analytics, and thus required further training on a particular software suite. The SAS Institute offered a variety of training programs for faculty. The SAS software, which included the SAS Enterprise Miner, including all the datasets used for the course was installed within the university's Citrix™ virtual server environment and made accessible to the students.

During the early discussions, the notion of going for a fully open-source platform for data analytics instruction was considered, but shelved in the end, as taking that route would require much more training, with not much support, and because many companies in our target areas may not be interested in open-source solutions for a variety of reasons. In addition, SAS offered a "Joint Certificate Program" for universities that joined the SAS Global Academic Program. This enabled the students who had taken four courses in which SAS software was used, and then completed a capstone project using SAS to acquire a certificate jointly signed by the university and SAS. This certification offered by SAS was a further incentive for the university to join the SAS Global Academic Program. Other incentives of the SAS Global Academic Program included: Free faculty training at the faculty workshops; 50% off the standard price on other training workshops; free use of SAS' instructional material for any of their courses, including access to the data files; two free SAS Press textbooks per faculty member per semester; and permission to use the SAS logo on the certification.

The course sequence that evolved after the initial discussions was as follows:

Required Core Courses (24 credits)		
Course	Title	Credits
BAN 610	Statistics and Probability	3
BAN 615	Predictive Modeling	3
CIS 620	Data Management	3
CIS 627	Data Warehousing	3
CIS 628	Data Mining	3
BAN 620	Text Mining	3
BAN 650	Data Visualization	3
BAN 690	Business Analytics Capstone	3
Elective Courses (12 credits)		
Students may select any 4 courses (12 credits) from the list below. Additional elective business courses are available to students at the discretion of the program director.		
BAN 660	Optimization	3
BAN 661	Web Analytics and Web Intelligence	3
BAN 662	Healthcare Informatics	3
BAN 663	Insurance Analytics	3
BAN 664	Enterprise Risk Management and Governance	3
CIS 625	ERP Design and Implementation	3
CIS 630	Business Design and Object-oriented Analysis	3
CIS 690	Managing Information Technology Projects	3

In our application for joint certification, we listed the following courses: BAN 610 (Statistics and Probability), BAN 615 (Predictive Modeling), BAN 620 (Text Mining), CIS 628 (Data Mining), and BAN 690 (capstone). We also described the courses and their learning outcomes then discussed how we were going to use SAS within each of the courses.

In the above, the core courses sequentially follow one another. The students thus acquire (or regain familiarity with) knowledge in the foundational aspects such as statistics and probability, as well as the basics of the SAS in BAN 610, followed by BAN 615. BAN 615 forms the basis of the two-part “data analysis” core. The topics covered in BAN 615 include: Data Preparation, Principal Components Analysis, Decision Trees, Neural Networks, Regression (including logistic and polynomial regression), Model comparison, assessment, scoring and implementation. The second part of the data analysis core is covered in CIS 628 (Data Mining). This course introduces the students to the overall concepts of data mining, and includes topics such as: Market Basket Analysis, Link Analysis, Survival Analysis, Genetic

Algorithms and Memory-Based Reasoning. Both of these courses are conducted using the SAS Enterprise Miner™ software. Between them, these two courses adequately prepare the student to get deeper into Text Mining, which is covered in BAN 620 (i.e. the course that follows BAN 610, BAN 615 and CIS 628).

TEXT MINING

The Internet and the World Wide Web, coupled with exponentially increasing computing power, has facilitated the explosion of social media, which in turn generates vast amounts of unstructured, textual data. Corporations and other organizations have been quick to realize the potential of mining this vast trove of data in order to identify, analyze, and categorize the sentiments of the participants of various social media in a variety of situations. The potential of text mining to analyze reactions, reviews, complaints, kudos, etc. has led to rapid growth in its usage in a variety of circumstances. Thus, text mining, even though it is a subset of the vast field of data mining, nevertheless occupies an important place in terms of its importance to BA education.

The course BAN 620 (Text Mining) was offered for the first time in the MSBA program in Fall 2014. During the preparations for the course, it was noticed that the SAS Enterprise Miner™ that was bought by the university did not come with the Text Miner module. The Text Miner module contained specialized features and modules that were tailored for text mining, and hence we made the decision to buy that module. It is important to note that SAS also offers a Sentiment Analysis Studio that is specialized towards performing text sentiment analysis. However, upon research, we concluded that a significant amount of sentiment analysis could be accomplished merely by doing a detailed analysis of the results from the Text Miner module, and hence decided against buying the Sentiment Analysis Studio. Further, it was also our conclusion that using the studio would preclude the students from doing their own detailed analysis of the text, and hence would not be very appropriate as a teaching tool for text mining.

TEXT MINING: COURSE DESIGN

In designing the course, we reviewed the SAS text mining materials. The instructor tasked with teaching the course enrolled in a web-based text mining course offered by SAS. We also reviewed the book “Text Mining and Analysis” by Chakraborty et al (Chakraborty et al, 2013), and consulted with Gautam Chakraborty and acquired some materials used in his text mining course at the Oklahoma State

University. Our syllabus had to take into account the online nature of the course as well as the short time duration, i.e., 7 weeks in all. Thus several adaptations were made to the materials offered by SAS and Chakraborty. The final list of topics was as follows:

- Introduction to text analysis
- Overview of text analytics
- Algorithmic and methodological considerations in text mining
- Applications of text mining to pattern discovery
- Applications of text mining to predictive modeling
- Introduction to text sentiment analysis
- Sentiment analysis of social media data – long project

It should be noted that even though sentiment analysis appears at the end of this list, the concept of ‘why analyze text?’ was introduced from the beginning of the course. Looking at the ‘why’ in this manner naturally led to sentiment analysis. The project’s general theme was introduced fairly early in the course. It was an open-ended exercise that was designed to focus the students’ attention to the various aspects of text mining and analysis while grounding those aspects on real data. A complete description of the project as well as the grading rubrics is given below.

Course Logistics and Student Composition

The course was conducted in an online format using the Blackboard Course Management System. The course consisted of a combination of online videos and lectures, reading materials, assigned discussions, lab assignments, and the final project. The students were working professionals from all over the country. Most were from the consulting and insurance industry, and had a fairly reasonable background in statistics. CIS628 (Data Mining) is a prerequisite for this course.

TEXT MINING: THE PROJECT

Project Description and Instructions

This is a very open-ended project in which you can do much (or less) analysis of textual data that has been presented to you.









Background

Two major announcements were made on November 10 and November 11, 2014. The first was a speech by President Obama, favoring Net Neutrality and urging the FCC (the US Federal Communications Commission) to favor the same in its

directives. The second was a joint announcement by President Obama and Chinese President Xi Jinping on their decision to fight climate change jointly. These announcements resulted in a lot of tweets in Twitter world. You have been given a Twitter dataset containing Tweets from the week prior to, and well as one week after the announcements.

The Twitter feeds I have acquired are in 9 XLS files. They were originally in CSV (comma separated values) format (each approx. 75 MB), but I converted them to XLS which resulted in reducing the size of each file to about 22 MB in size. Each file contains about 98,000 rows (records or tweets).

Upon looking at the files, I realized that I had to “prepare the data.” So you also have to edit the file (the one you choose to work on) headers so that the column titles would make sense. That is file “11-03 to 11-05 twitter data1.xls”. The other files are:

	11-05 to 11-09 twitter_data2.xlsx	11/13/2015 4:50 PM	Microsoft Excel W...	22,121 KB
	11-09 to 11-10 twitter_data3.xlsx	11/13/2015 4:48 PM	Microsoft Excel W...	21,448 KB
	11-10 to 11-11 twitter_data4.xlsx	11/13/2015 4:46 PM	Microsoft Excel W...	21,670 KB
	11-11 to 11-12 twitter_data5.xlsx	11/13/2015 4:44 PM	Microsoft Excel W...	21,922 KB
	11-12 to 11-13 twitter_data6.xlsx	11/13/2015 4:42 PM	Microsoft Excel W...	22,487 KB
	11-13 to 11-15 twitter_data7.xlsx	11/13/2015 4:40 PM	Microsoft Excel W...	22,455 KB
	11-15 to 11-17 twitter_data8.xlsx	11/13/2015 4:38 PM	Microsoft Excel W...	21,472 KB
	11-17 to 11-17 twitter_data9.xlsx	11/13/2015 4:36 PM	Microsoft Excel W...	9,590 KB

1. Data Preparation: You need to first prepare the data. For example, if you open up 11-03 to 11-05 twitter data1.xls, you will see that the column titles are not very clear. You may have to edit the existing titles in order to clean them up. You may also have to remove some empty columns.
2. I suggest that you use only any one file at first. After preparing the XLS file, you have to import it into Text Miner, and then do text analysis.
3. Note that you are analyzing the actual Text in the Tweets. Now, I noticed something when I loaded and worked on the data file: SAS Text Miner creates Metadata in such a manner that the longest variable is denoted as the “Text” variable, and only that variable’s contents are analyzed!! That will not work for us, because you will notice there are other column variables that have strings that are longer than 140 characters! So you need to figure out how to make SAS Text Miner designate the actual Tweets are the real text for this exercise.

4. Later on, you can decide if you want to combine two files (e.g. one pre- and one post-announcement) and see if there is any qualitative change in the reactions (via the tweets) or to find new clusters if any and analyze them.
5. Please do not try to add all the given files! You may crash the Text Miner or worse, hang Citrix!
6. I will be quite satisfied if you use just 2 files in this project.
7. Your work: This is where your creativity comes in! At a minimum, you could identify clusters among the tweets. Then you can try to increase or reduce the clusters, if any. (So this would be undirected Text mining)
8. Hints: Obviously, you have to parse, filter, etc., etc. You may find the Demos in Mod 4 to be useful templates.
9. You can then consider loading the clusters into a Decision Tree and see if you can use additional data (a subset of one of the above files) to “Score” the data and put the data into clusters.
10. More importantly, you need to start with your goals for the project, and then discuss in detail how you arrive at the goals, what methods you used and why, and what you have learned about the data itself from your analysis.

As you can see, this is a very open-ended project, (and very real-life) where you use available text data and use your creativity to identify/analyze sentiments and patterns!

Deliverables

You will demonstrate your knowledge of Text Mining using the SAS Enterprise Miner’s Text Miner Module. You will analyze the textual data and answer some specific questions posed on:

1. The goal(s) of the project
2. The method used and the reasoning behind the method used to achieve the goal(s)
3. An analysis of the solution(s) obtained
4. Problem/goal extension

Grading Rubrics

The grading rubrics basically followed the Assessment of Learning (AOL) for this course. Assessment of Learning focused on the students’ ability to choose the appropriate analytic method(s) to analyze the data, the actual analysis of the data, perform an appropriate interpretation of the analysis done on the data, and finally, demonstrate how the information gained in this project could be used to push

beyond the scope of this project. The students' work was then graded based on a points-based system given below.

I. Analytic Method

- _____ 3. Superior: Accurately and correctly uses appropriate methods for solving the problem.
- _____ 2. Acceptable: Few errors in the methods used for solving the problem.
- _____ 1. Needs Improvement: Exhibits insufficient knowledge of methods for solving the problem.

II. Interpretation of Analysis

- _____ 3. Superior: Thoroughly describes the information gained with clear observations from doing the project.
- _____ 2. Acceptable: Adequately describes information gained from the project.
- _____ 1. Needs Improvement: Observations about information gained are underdeveloped.

III. Problem Extension

- _____ 3. Superior: Goes over and beyond scope of the project. Information gained adds to the knowledge base for this topic.
- _____ 2. Acceptable: Provides adequate information to satisfy the knowledge base for this topic.
- _____ 1. Needs Improvement: Problem extension is underdeveloped or non-existent.

In addition to the above, I will also score the goals you define for your project.

TEXT MINING: PROJECT REPORT

With the above as background, we then proceed to look at the deliverable that came out of the course. We illustrate that by using a representative sample of a report that was produced by a team of two students. As seen in the 'rubrics' above, the report was required to document the analytic method, interpretation and problem extension that sought to illustrate the process of text analysis. The original report has only been minimally edited, so as to preserve the actual 'voice' and thought process of the students as they went about this exercise. This particular project was selected because it consistently placed in the "superior" category in each of the evaluation points set forth in the rubrics above.

The representative project is titled "Net Neutrality through Twitter and Text Mining." This particular team thus chose to focus on the net neutrality component of the Twitter data that was given to them. The report is given in the Appendix. The

report has some interesting features that can be used as a pedagogical template in similar courses. The report provides a step-by-step approach to using SAS to analyze Twitter data, starting from data preparation. After the raw data is prepared, it is imported into SAS Enterprise Miner. After that, the data is processed through various nodes which include both the generic Enterprise Miner nodes as well as the specialized Text Miner nodes. The reasoning behind the use of each node is explained. The nodes through which the data is processed include the Text Parsing Node, Text Filter Node, Text Cluster Node and Text Topic Node. The reasoning and justification that are used in selecting certain terms for clustering or a more detailed analysis is discussed. Finally, a detailed interpretation of the text analysis is provided, along with how this type of model could be extended to other problems. The report also includes a complete set of references at the end which provides more details of the Twitter data sources as well as cross references between Twitter data and news articles in other media.

Overall, the report provides a complete and detailed picture of the entire process of text mining.

CONCLUSION

The field of data analytics has gained tremendous momentum in the past five years. This has greatly been aided by advanced computing tools for data analysis, as well as the virtual tsunami of data that is generated through social media applications as well as other routinized transactional data collection methods. Textual data has become an important aspect of data analytics in a variety of disparate fields such as retail business analysis, consumer satisfaction studies, politics, and public opinion and public policy spheres. The growth in analytics has created a tremendous need for developing university curricula in the field. It can be conjectured that the greatest need for advanced education in this field exists for those who are already working in this field, and who may already have a basic degree, so as to introduce them to new tools and techniques to enhance their knowledge. Thus the need exists for on-line delivery of such programs using advanced course delivery tools.

In this paper we discuss the development of an online graduate program in business analytics, and then delve deeper into the development of a specific course, namely Text Mining. We describe the rationale behind the system and software choices that were made in developing the program. Then we provide a detailed description of the Text Mining course, along with a key project that was required to be completed in the course. A sample project is provided as a template that other faculty could possibly use as a teaching tool or to fashion their own assignment or project. We

show through the project how the text analysis of Twitter data can be analyzed for sentiments, without having to use the SAS Sentiment Analysis Studio.

Overall, the course was a success. The student-participants expressed satisfaction in the course content and the project in particular. Several remarked that their projects would serve as templates for similar projects in their work environments. For the instructor, this exercise proved that social media reactions to local current events can successfully be used as a learning tool in a course in text mining. The use of Twitter and current events served to increase the participants' interest in the exercise. The use of SAS, especially the "Term-editor" allowed the student-participants to view the terms that were selected for clustering by SAS. They could modify or remove the terms, or increase the emphasis or important of certain words or phrases. This feature is important, as it enables the students to think carefully about various possibilities and combinations of words and phrases and their effects on clustering and sentiment analysis. In a future course, we feel that we could include a comparative aspect to the exercise. For example, sentiments could be categorized by running the tweets through the "OpinionFinder" open source software and compare the results with those that were arrived at using SAS. An added exercise would be to run a regression on the results of the OpinionFinder using R, another open-source software. This will enable students to compare the features of SAS, which is proprietary software, with open source software.

REFERENCES

Chakraborty, G, Murali Pagolu, Satish Garla (2013). Text Mining and Analysis: Practical Methods, Examples and Case Studies Using SAS®. SAS Institute, Cary, NC, USA. P320.

APPENDIX

Report: Net Neutrality Through Twitter and Text Mining (Student Co-Authors: Danielle Cote and Jacelyn Locke)*

(*The students who wrote this report are co-authors of this paper, and as such, have permitted the presentation of this report.)

INTRODUCTION

In November of 2014, President Barack Obama addressed the nation by calling on the Federal Communication Commission (FCC) to take up the strongest possible rules to protect net neutrality (Mechaber, 2014). In this two-minute internet video hosted on YouTube, “President Obama’s Statement on Keeping the Internet Open and Free”, Obama states “The internet has become an essential part of communication and everyday life... I am asking the FCC to make sure that the consumers, not the cable company, gets to decide which sites they use.” According to the definition of net neutrality on The White House’s webpage, net neutrality is the idea that all internet traffic should be created equally; “an entrepreneur’s fledgling company should have the same chance to succeed as established corporations, and that access to a high school student’s blog shouldn’t be unfairly slowed down to make way for advertisers with more money.” Following the speech, Americans and companies reacted through various internet sites, including the social media site Twitter. In order to understand the public’s reaction to Obama’s speech, text mining can be used to hash through the tweets following the speech. Text mining is a machine learning process that enables researchers to extract high quality information to draw conclusions from a vast amount of text. With the use SAS Enterprise Miner, a data and text mining tool, the text of thousands of tweets can be manipulated and then clustered to determine the overall public reaction.

BACKGROUND

During his campaign in 2007, Senator Barack Obama pledged Net Neutrality laws if elected President of the United States. Obama made statements against lobbyist in favor of being able to be gatekeepers and to charge different rates to different Web sites, expressing that websites such as Google may not have gotten to where they are today without a level playing field (Broache, 2007). His campaign urged the freedom and equality of the internet. As promised, by 2010 the FCC passed their

final laws to protect net neutrality by requiring “Internet service providers to give consumers equal access to all lawful content without restrictions or tiered charges” (Cameron, 2013).

Shortly after the FCC passed their first ever rules to protect net neutrality, Comcast Communications, better known as Verizon, filed a federal lawsuit against them. Verizon argued “that the FCC does not have the legal authority to mandate how Internet service providers treat content on their networks” (Kang, 2011). Michael E. Glover, Verizon's senior vice president and deputy general counsel stated, "We believe this assertion of authority goes well beyond any authority provided by Congress, and creates uncertainty for the communications industry, innovators, investors and consumers" (Kang, 2011). On January 14, 2014 in *Comcast Corp. v. FCC*, the United States Court of Appeals for the District of Columbia ruled that the FCC does not have jurisdiction of Comcast's Internet service under the language of the Communications Act of 1934. The Communications Act of 1934, signed into law by President Franklin D. Roosevelt, created the FCC to oversee and regulate industries such as telephone, telegraph, radio, broadcast, and television communications. Judge David Tatel stated for a three judge panel, "Given that the Commission has chosen to classify broadband providers in a manner that exempts them from treatment as common carriers, the Communications Act expressly prohibits the commission from nonetheless regulating them as such."

A day after the ruling, a petition was created, and then signed by 105,572 users, on the White House platform for the Obama administration to "Restore Net Neutrality by directing the FCC to classify Internet providers as 'Common Carriers'." The White House eventually responded to the petition in agreement but stated that it was unable to direct the FCC rulemaking. FCC decided to open a four-month window from May to September of 2014 for anybody to leave comments regarding their opinion on net neutrality. The FCC received a record breaking of 3.7 million comments on this issue (Hu, 2014). It was after this window that the White House posted the video, “President Obama's Statement on Keeping the Internet Open and Free”, on YouTube.

PROBLEM STATEMENT

After the announcement mentioned above many people went online to various social media sites to post their responses to the video. One of the sites used was Twitter and thousands of “tweets” were generated around the time of the announcement about Net Neutrality. From these tweets we should be able to get an idea on the public's opinion on Net Neutrality and their response to the president's

announcement. However, these tweets are considered unstructured data and can be difficult and very time consuming to analyze. In order to obtain this information and gather details about the general opinion of Net Neutrality in the US we would need to gather this unstructured data from Twitter and use a specialized tool to analyze the tweets. The SAS Enterprise Miner is a data and text mining tool that will be able to take the unstructured data from the tweets and generated useful information. We can then use this information to determine what the public's opinion on Net Neutrality was before the announcement and what their general opinions were of President Obama's announcement. This information could be very useful to the White House to get a list of reactions to the President's announcement and help to determine next steps in order to either address certain reactions or adjust future announcements in order to get reactions closer to what they would want.

ANALYTIC METHODS

Figure 1: Analyzing the 'Day Before Announcement' Data

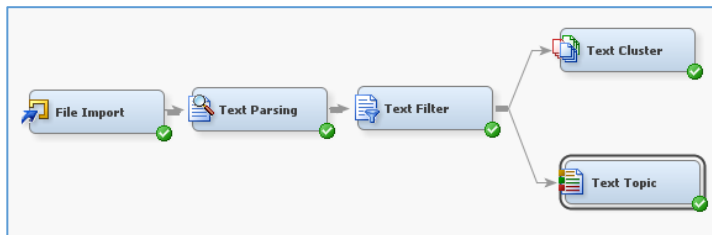


Figure 1 shows the diagram used in this analysis. The text file used for the “Day Before” part of the analysis was “11-05 to 11-09 twitter_data2.xlsx” from a few days before Obama’s announcement on Net Neutrality to the day right before the announcement. Before the file could be imported using SAS Enterprise Miner, there was a need to eliminate some of the columns from the data set as they added no additional value to the analysis. There was also a need to change the names of the columns in order for a proper descriptive column name to be able to transfer successfully over to the tool. After these changes were made to the data set was ready to be imported into the SAS Enterprise Miner tool.

Figure 2

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Country	Input	Nominal	No		No	.	.
Hashtag	Text	Nominal	No		No	.	.
Link	Input	Nominal	No		No	.	.
Location	Rejected	Nominal	No		No	.	.
Text	Text	Nominal	No		No	.	.
Time	Time ID	Interval	No		No	.	.
Title	Input	Nominal	No		No	.	.
URL	Rejected	Nominal	No		No	.	.

The first node added to the diagram was the File Import node. This node was added in order to import the excel file into SAS Enterprise Miner and to assign roles to the variables. The Hashtag column was also given the Text Role as this column held terms that would also be valuable for this analysis. The URL and Location variable were given Rejected Roles as they would not be needed for this analysis. Figure 2 shows the variables roles and levels used in the analysis.

The next node to be added is the Text Parsing node. This node was added to help quantify the terms found in the various tweets and to see if certain terms were used more than others. This node was used with its default property settings. The next node to be added to the diagram was the Text Filter node. This node was added to help filter out tweets that did not pertain to the net neutrality topic. In Figure 3 you can see the terms “Neutrality” and “Net” were added along with their synonyms to the Interactive Filter Viewer in order to analyze just the tweets about Net Neutrality.

Figure 3

Interactive Filter Viewer

File Edit View Window

Search: >#neutrality net Apply Clear

Documents

TEXT	TEXTFILTER2_SNIPPET	TEXTFILTER2_RELEVANCE	COUNT
Verizon appears to soften its stance against net neutrality rules http://t.co/jvCFha6G9GK	... its stance against net neutrality	0.5	
Net neutrality was the biggest tech issue of the year. But nobody campaigned on it.	... Net neutrality was the biggest	0.5	
RT @SteveForbesCEO: No matter how hard the left pushes net neutrality, voters see through	... the left pushes net neutrality ,	0.5	
RT @EFF: An FCC "hybrid" proposal that grants net neutrality rules for Internet companies but	... proposal that grants net	0.5	
My boss Gets It! Net Neutrality is Essential to Growing Your Business and Brand	... Gets It! Net Neutrality is	0.5	
#SocialMedia Net Neutrality is Essential to Growing Your Business and Brand	... # SocialMedia Net Neutrality is	0.5	
emergency net neutrality vigil tomorrow in Boston at 6pm EST in front of the state house	... emergency net neutrality vigil	0.5	
RT @SteveForbesCEO: No matter how hard the left pushes net neutrality, voters see through	... the left pushes net neutrality ,	0.5	
#ToucheComm : Net Neutrality is Essential to Growing Your Business and Brand: Video has	... # ToucheComm : Net Neutrality	0.5	
#Entrepreneur Net Neutrality is Essential to Growing Your Business and Brand	... # Entrepreneur Net Neutrality is	0.5	
Net Neutrality is Essential to Growing Your Business and Brand http://t.co/Tavay379PV	... Net Neutrality is Essential to	0.5	
New Find Net Neutrality is Essential to Growing Your Business and Brand	... New Find Net Neutrality is	0.5	
Net Neutrality is Essential to Growing Your Business and Brand http://t.co/jubayENWBU9	... Net Neutrality is Essential to	0.5	
Net Neutrality is Essential to #FreeSpeech & #Democracy	... Net Neutrality is Essential to #	0.5	

Terms

TERM	FREQ	# DOCS	KEEP	WEIGHT	ROLE	ATTRIBUTE
neutrality	20548	9105	<input checked="" type="checkbox"/>	0.014	Noun	Alpha
protest	5472	4062	<input checked="" type="checkbox"/>	0.101	Noun	Alpha
rt	3838	3794	<input checked="" type="checkbox"/>	0.103	Prop	Alpha
fcc	3054	2950	<input checked="" type="checkbox"/>	0.131	Prop	Alpha
city	2066	2060	<input checked="" type="checkbox"/>	0.169	Noun	Alpha
net	1872	1863	<input checked="" type="checkbox"/>	0.18	Prop	Alpha
tonight	1997	1092	<input checked="" type="checkbox"/>	0.241	Noun	Alpha
plan	1151	1086	<input checked="" type="checkbox"/>	0.241	Verb	Alpha
show	891	891	<input checked="" type="checkbox"/>	0.26	Verb	Alpha
map	890	890	<input checked="" type="checkbox"/>	0.26	Verb	Alpha
neutrality	817	814	<input checked="" type="checkbox"/>	0.27	Prop	Alpha
hybrid	814	813	<input checked="" type="checkbox"/>	0.27	Noun	Alpha

The next node to be added to the diagram was the Text Topic node. This is very useful to the analysis because it is able to form “topics” by grouping together associating terms. By using this node we are able to get an idea about the major themes or topics among the numerous tweets. After running the Text Topic node with its default settings we made the decision to lessen the number of topics to 10 since the original 25 showed overlapping themes. Figure 4 shows the properties for the Text Topic Node.

Figure 4

Property	Value
General	
Node ID	TextTopic2
Imported Data	
Exported Data	
Notes	
Train	
Variables	
User Topics	
<input checked="" type="checkbox"/> Term Topics	
Number of Single-term Topics	0
<input checked="" type="checkbox"/> Learned Topics	
Number of Multi-term Topics	10
Correlated Topics	No
<input checked="" type="checkbox"/> Results	
Topic Viewer	

Figure 5

Property	Value
General	
Node ID	TextCluster2
Imported Data	
Exported Data	
Notes	
Train	
Variables	
<input checked="" type="checkbox"/> Transform	
SVD Resolution	Low
Max SVD Dimensions	100
<input checked="" type="checkbox"/> Cluster	
Exact or Maximum Number	Maximum
Number of Clusters	10
Cluster Algorithm	Expectation-Maximization
Descriptive Terms	10

The node added to the diagram next was the Text Cluster Node. This node was added because it is able to group documents together based on their descriptive terms. This will be very useful for our analysis especially since this node also reports out these descriptive terms as well as some details pertaining to them. We decided to change the maximum number of clusters to 10 since that was the number used in the Text Topic node. After running the node once we decided to decrease the number of descriptive terms to 10 since there seems to be repeating terms. Figure 5 shows the properties for the Text Cluster Node.

Analyzing the ‘Day After Announcement’ Data

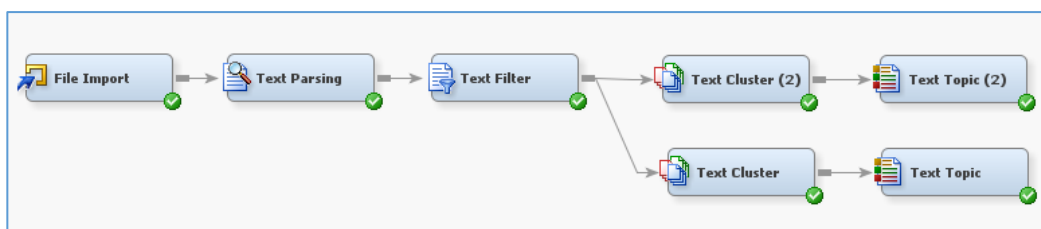
Figure 6

Figure 6 shows the diagram used in this part of the analysis. The text file used was “11-11 to 11-12 twitter_data5.xlsx”, data from 1-2 days after Obama’s announcement on Net Neutrality. Before modeling began in SAS Enterprise Miner, the column names of this data was reformatted and the column containing all blanks labeled “UnitId” was removed. The file was imported into SAS Enterprise Miner using the File Import node. Some of the text variables were rejected in order to do proper analysis of the tweets (Figure 7). A Text Parsing node with default settings

was then attached to the File Import Node. Next, a Text Filter node was added. The Term Weight property was changed to Entropy because there is no target variable. Lastly, a Text Cluster and Text Topic node were attached with default settings. Using these default settings, the clusters in Figure 8 were formed.

As seen in Figure 8, most of the clusters deal with Net Neutrality and Climate Change. Since we are only concerned with Net Neutrality, utilizing the Filter Viewer in the Text Filter node will eliminate the Climate Change clusters and any other non-relevant tweets. In the Filter Viewer, we filtered on “+net”. Also, changing the default settings of the Text Cluster and Text Topic node to a smaller amount of clusters will eliminate the chance of repeating tweets. Another Text Cluster Node was added to the diagram and the default settings of Number of Clusters was changed from 40 to 10. The Descriptive Terms defaults settings were also changed from 15 to 8. After running the Text Cluster node, 6 clusters were formed. Next, a Text Topic node was attached to the Text Cluster node. The default settings on this node was also changed. The ‘Number of Multi-term Topics’ was changed from 25 to 10.

Figure 7

Variables - FIMPORT

(none) ☐ not Equal to

Columns: ☐ Label ☐ Mining ☐ Basic

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
country_code	Input	Nominal	No		No	.	.
favorites_count	Input	Interval	No		No	.	.
followers_count	Input	Interval	No		No	.	.
friends_count	Input	Interval	No		No	.	.
hashtag	Text	Nominal	No		No	.	.
id	ID	Nominal	No		No	.	.
is_retweet	Input	Nominal	No		No	.	.
link	Rejected	Nominal	No		No	.	.
location_coords	Rejected	Nominal	No		No	.	.
location_coord_t	Input	Nominal	No		No	.	.
location_display	Input	Nominal	No		No	.	.
location_type	Input	Nominal	No		No	.	.
media_display_u	Rejected	Nominal	No		No	.	.
media_type	Input	Nominal	No		No	.	.
media_url	Rejected	Nominal	No		No	.	.
posted_time	Time ID	Interval	No		No	.	.
real_name	Input	Nominal	No		No	.	.
source	Input	Nominal	No		No	.	.
statuses_count	Input	Interval	No		No	.	.
Text	Text	Nominal	No		No	.	.
Title	Text	Nominal	No		No	.	.
tweet_url	Rejected	Nominal	No		No	.	.
username	Input	Nominal	No		No	.	.
user_bio_summa	Rejected	Nominal	No		No	.	.
user_location	Rejected	Nominal	No		No	.	.
user_mention	Rejected	Nominal	No		No	.	.
user_mention_u	Rejected	Nominal	No		No	.	.
user_twitter_pa	Input	Nominal	No		No	.	.

Figure 8

Clusters			
Cluster ID	Descriptive Terms	Frequency	Percentage
1	net neutrality +debate 'net neutrality' obama 'net neutrality debate' neutral +fall access top costs issues divisive 'divisive debate' +pres...	2202	2%
2	climate china deal reach +talk secret +secret talk haunting +photo beautiful polar +beautiful polar photo +haunt story +story +wire ...	5611	6%
3	neutrality net 'net neutrality' +vote +know +delay +care +good n't +support +conservative access https +company +rule	6007	6%
4	neutrality net obama 'net neutrality' internet +push obamacare lbd fcc +president rules http://t.co/wdfidem6bo http://t.co/ey12smxro +up...	4740	5%
5	climatechange it amp +action auspol +energy +target +announcement usa global abbott historic leadership globalwarming +climate...	5778	6%
6	neutrality net comcast 'net neutrality' +statement +fix bogus misleading +agree +president 'misleading net neutrality statement' +want g...	2518	3%
7	netneutrality internet it open +free +keep barackobama +want ofa retweet +sign +petition +gatekeeper +agree +add	5274	5%
8	climate china +deal obama +agreement +emission +cut historic +announce +reach 'climate change' climatechange +president gree...	8559	9%
9	netneutrality barackobama internet obama title ii tomwheelerfcc fcc +free open +keep +president +plan amp +want	5606	6%
10	work 'hybrid approach' +martial art movie fight +approach +art +fight +minute +movie +scene +work citizenradio http://t.co/y32cdfnu...	650	1%
11	netneutrality obama internet http fcc +president https amp +want +protect +support +victory +win huge +stand	8302	8%
12	neutrality net fcc obama 'net neutrality' +plan wheeler 'obama's net neutrality plan' tom chairman +ignore +chairman +chair +reject +w...	4281	4%
13	climate 'climate change' +world +scientist +study dead +prepare +zone +country +good republicans earth +fight +action +look	5701	6%
14	amp climate +lobbyist +appoint +run +protester +debate +sing +net neutrality protester +start telecom monday +end +cable tom	6312	6%
15	climate it 'climate change' china http abbott amp +deal +big auspol +action +reality +world polar +photo	6471	7%
16	20 +climate climatechange +agenda onmyagenda auspol 'climate change' australia abbott summit tonyabbottmhr otiose94 http://t.co/...	2328	2%
17	neutrality net internet 'net neutrality' obamacare cruz ted sentedcruz oatmeal giants push +sideline +government +speed +explain	8392	8%
18	china +climate ambitious +unveil +goal +agree 'climate change' +deal +emission goals unveil +ambitious climate change goalf +clim...	4465	5%
19	ballmer net neutrality steve total +total tweet dismiss' dismissing spends tweets techcrunch fan plan tech 'obama's plan' chalk	589	1%
20	neutrality net cruz ted oatmeal 'net neutrality' +explain +work dear senator +backfire ignoramus tweet bretzysbs sentedcruz	5224	5%

INTERPRETATION

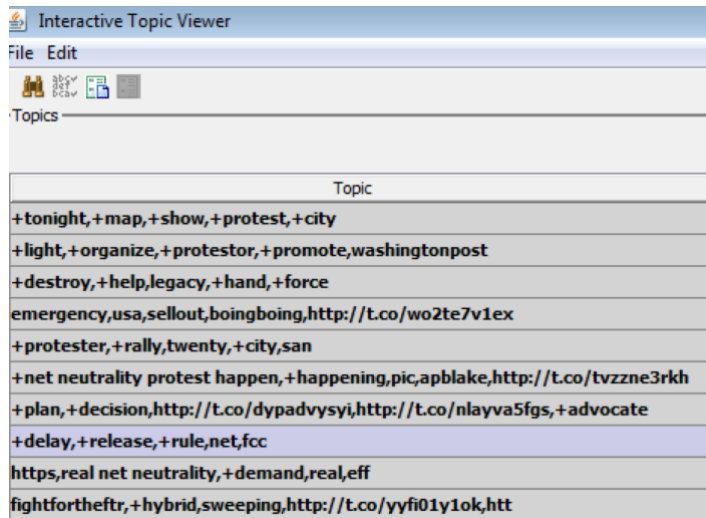
‘Day before’

After running the diagram to the Text Cluster node the SAS Enterprise Miner tool was able to generate 6 clusters. The first cluster contains the terms “fcc +rule net +delay +release Verizon hybrid +sue +approach +work”. Similar terms were found in the Text Topic node in one of the topics generated by the diagram. This topic included the terms “+delay, +release, +rule, net, fcc”. Using the Interactive Topic viewer we were able to see the tweets that included these terms. The clusters and topics generated are shown in figures 9 and 10, respectively.

Figure 9

Clusters			
Cluster ID	Descriptive Terms	Frequency	Percentage
1	fcc +rule net +delay +release verizon hybrid +sue +approach +work	1043	11%
2	city +protest +tonight +map +show +house white http +rally +protestor	1889	19%
3	protest rt fcc +plan emergency +hybrid +advocate usa internet +decision	2583	27%
4	neutrality +rally +protester +city net +destroy +help obama legacy +force	869	9%
5	https +demand real +join eff today netneutrality 'real net neutrality' +net +emer...	2177	22%
6	neutrality net isp tech throttling essential growing +brand +balloon +trial	1144	12%

Figure 10



Topic
+tonight,+map,+show,+protest,+city
+light,+organize,+protestor,+promote,washingtonpost
+destroy,+help,legacy,+hand,+force
emergency,usa,sellout,boingboing,http://t.co/wo2te7v1ex
+protester,+rally,twenty,+city,san
+net neutrality protest happen,+happening,pic,apblake,http://t.co/tvzzne3rkh
+plan,+decision,http://t.co/dypadvysyi,http://t.co/nlayva5fgs,+advocate
+delay,+release,+rule,net,fcc
https,real net neutrality,+demand,real,eff
fightforthefttr,+hybrid,sweeping,http://t.co/yyfi01y1ok,htt

Figure 11

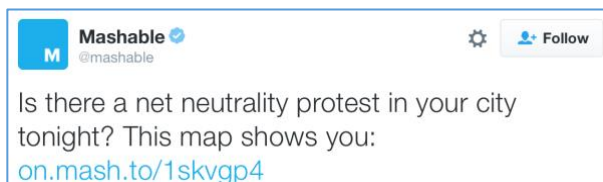


We will discuss the tweets pertaining to the top three clusters here. The tweet shown above (Figure 11) was made by CNET and comments on the FCC's delay in releasing rules for Net Neutrality with a link to one of their articles detailing the issue and explaining what Net Neutrality is. This post was retweeted over 300 times which is why is generated both a topic and cluster in the diagram used for this analysis. These tweets show that many were concerned about the decision that the FCC was going to make about Net Neutrality and were keeping others updated on this decision making progress.

The second cluster in the Text Cluster node contains the terms "+city +protest +tonight +map +show +house white http +rally +protestor". This cluster can

be matched to the topic that contains the terms “+tonight, +map, +show, +protest, +city” and the topic that contains the terms “+light, +organize, +protestor, +promote, Washington post”. By using the Interactive Topic viewer we can see that the first topic was also generated from numerous retweets. The original tweet is given in Figure 12.

Figure 12



This tweet (Figure 12) was tweeted over 900 times and gives insight into the public's opinion on the Net Neutrality topic. The protests mentioned in this tweet are for Net Neutrality meaning that most do not want any company or government body putting limits on the content able on the internet. It seems that there were many protest held in the US over Net Neutrality and Mashable created an interactive map to show users where the nearest protest was to them. People retweeted this post in order to get others to come and protest and essentially fight for Net Neutrality. The second topic was generated by tweets and retweets of the Washington Post (Figure 13).

Figure 3



This post was tweeted by the Washington Post in order to share one of their articles that details a small protest held outside the White House to promote Net Neutrality. This protest seemed to be in response to people believing that the FCC was considering a plan that would give some power over Net Neutrality over to the internet providers. This tweet was shared over 300 times and also shows that many are in favor of Net Neutrality and do not want internet providers to have any control over it.

The third cluster in the Text Cluster node contains the terms “+protest rt fcc +plan emergency +hybrid +advocate usa internet +decision” (Figure 14). This cluster can be matched to the topic that contains the terms “emergency, usa, sellout, boingboing, <http://t.co/wo2te7vlex>” and the topic that contains the terms “+plan,+decision,<http://t.co/dypadvysyi>,<http://t.co/nlayva5fgs>,+advocate”.

Figure 14



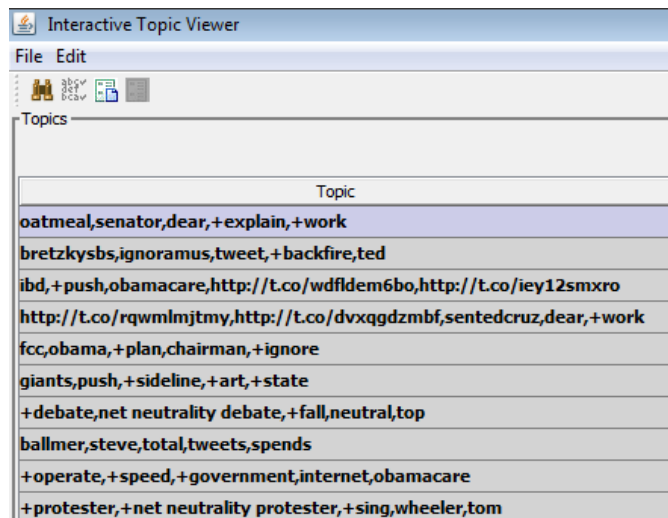
‘Day After’

Figures 15 and 16 show the clusters formed, and the topics generated in the SAS Enterprise Miner tool using data from the day after the announcement. The top two clusters and topics are discussed here.

Figure 15

Results - Node: Text Cluster Diagram: Final			
File Edit View Window			
Clusters			
Cluster ID	Descriptive Terms	Frequency	Percentage
1	comcast +statement +fix bogus misleading 'misleading net neutrality statement' +house white	1675	4%
2	'net neutrality' internet +vote +know +delay +good +rule +care	7563	20%
3	obama +president fcc +support 'obama's plan' +plan +conservative +reject	3065	8%
4	obama internet obamacare +push ibd http://t.co/wdtdem6bo http://t.co/ley12smro +lobbyist	2264	6%
5	net obama +debate +president rules 'net neutrality debate' netneutrality neutral	4917	13%
6	obama fcc +plan 'obama's net neutrality plan' wheeler chairman tom +ignore	4240	11%
7	internet rt obamacare cruz ted sentedcruz +debate push	5358	14%
8	cruz ted 'net neutrality' +explain +work dear oatmeal senator	5609	15%
9	oatmeal addresses +net verizon title ii +issue techcrunch	3570	9%

Figure 16

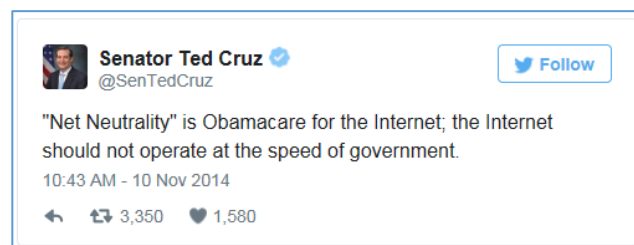


The screenshot shows the 'Interactive Topic Viewer' window. It has a menu bar with 'File' and 'Edit'. Below the menu bar is a toolbar with icons for file operations. The main area is titled 'Topics' and contains a table with a single column labeled 'Topic'. The table lists several topics, with the first one highlighted in blue.

Topic
oatmeal,senator,dear,+explain,+work
bretzkysbs,ignoramus,tweet,+backfire,ted
ibd,+push,obamacare,http://t.co/wdflde6bo,http://t.co/iey12smxro
http://t.co/rqwmlmjtmty,http://t.co/dvxqgdzmbf,sentedcruz,dear,+work
fcc,obama,+plan,chairman,+ignore
giants,push,+sideline,+art,+state
+debate,net neutrality debate,+fall,neutral,top
ballmer,steve,total,tweets,spends
+operate,+speed,+government,internet,obamacare
+protester,+net neutrality protester,+sing,wheeler,tom

The first topic contains 3,498 documents using the topics “oatmeal, senator, dear, +explain, +work”. This is also a series of retweets with the text “Dear Senator Ted Cruz, I’m going to explain to you how Net Neutrality ACTUALLY works - The Oatmeal http://theoatmeal.com/blog/net_neutrality ... via @Oatmeal”. The link brings you to The Oatmeal’s blog where it responds to Ted Cruz’s tweet in Figure 17.

Figure 17

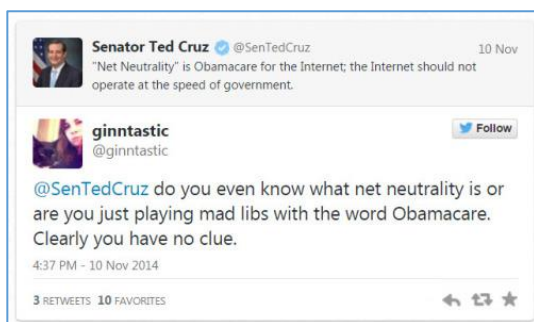


The blog post supports Net Neutrality and explains to Ted Cruz why his tweet does not make sense. This cluster of retweets also supports Net Neutrality. The fourth topic containing the words “<http://t.co/rqwmlmjtmty>,<http://t.co/dvxqgdzmbf>,sentedcruz,dear,+work” also shows the same retweet with 1,916 documents. However, the ninth topic contains the words “+operate, +speed, +government, internet, obamacare” and 1,585 documents were found. This topic is a retweet of Ted Cruz’s tweet, and shows

public support of Cruz's opinion. The two sides of Net Neutrality really start to appear in the analysis of Net Neutrality.

The second text topic is also a response to Senator Ted Cruz. The topic contains the following words "bretzkysbs,ignoramus,tweet,+backfire,tet" and 1,598 documents were found. The retweet is "bretzkysbs: Ted Cruz's Net Neutrality Tweet Backfires Ignoramus gets roasted on Twitter." The link the retweet brings you to bretzkysbs twitter page containing a series of tweets in reaction to Ted Cruz's tweet relating Net Neutrality to Obamacare. The tweets reacted to Ted Cruz's tweet show a strong movement towards Net Neutrality and irritation that Cruz compared Net Neutrality to Obamacare. A sample is given in Figure 18.

Figure 18



The next text topic contains the words "ibd, +push, obamacare, <http://t.co/wdflde6bo>, <http://t.co/iey12smxro>" and 1,479 documents were found. The retweet states "(IBD) Obama Pushes Net Neutrality For Control Of Web - 'ObamaCare Of Internet.'" This tweet on twitter shows a strong opposition to Obama's speech on Net Neutrality. Figure 19 shows the tweet along with a picture of Obama with the words "For your safety the internet must be controlled." There is also a link in the tweet bringing you to Investors Business Daily Politics webpage to an editorial named "Obama Pushes Net Neutrality For Control Of Web." This article is against what Obama stated in his speech and compares the control of the web to how China runs their country. The article states, "Obama wants the FCC to unilaterally put the Internet under heavily regulated Title II, which would apply 1934 Telecom Act landline law to the formerly unfettered and free Internet, in essence making the Web a government utility. Republican Texas Sen. Ted Cruz has justifiably called the president's proposal "ObamaCare for the Internet" (2014).

Figure 19

DISCUSSION

From the analysis on the day before data set it is clear that many people are in favor of Net Neutrality and there is a great concern that the internet may be limited in some way due to either a decision by the FCC or the government. From these tweets we can tell that many protests are formed for Net Neutrality right before a decision is made in order to impact the decision and have a more favorable outcome for the public. These protests are also popular right after a decision is made in hopes of changing the outcome. The tweets also showed how much people did not want to compromise on Net Neutrality due to the negative tweets about the hybrid plan to allow internet providers to control speed on some websites and also by the tweets pertaining to protesting for “real” Net Neutrality. This “real” Net Neutrality means that there are no limits to the internet and internet providers do not have any control on internet content or speed. From this analysis we can conclude that before the announcement by President Obama many were in favor of Net Neutrality and were hoping to persuade the president through protest to make decisions in favor of Net Neutrality. We can then predict that since the announcement by the president was in favor of Net Neutrality that many would have a positive reaction to his speech.

From the analysis on the day after data set we can conclude that the general public is in favor of a free internet, but are separated on how America should go about having a free internet. After Obama's speech there was a side of opinions in support of him, and others who agreed with Ted Cruz when he tweeted Obamacare for the internet. Obama's speech ignited many shares to websites that were strictly informative on Net Neutrality and did not necessarily show a side or try to persuade. These shared tweets are a positive sign that the general public is gaining more interest on the topic. The more interest gained on the topic will be essential in making a free internet the way the citizens of American want it.

In doing this analysis it was easy to see the impact that retweets had on the conclusions drawn and the topics and clusters generated in SAS Enterprise Miner tool. Since these retweets were so numerous other tweets ended up buried in a cluster or topic. It would be interesting to redo this analysis by either limiting the number of retweets or eliminating them all together. It would be interesting to see if any more themes or topics would emerge in this new analysis.

REFERENCES (FOR THE REPORT)

Broache, A. (2007, October 29). Obama pledges Net neutrality laws if elected president. CNET.

Retrieved from <http://www.cnet.com/news/obama-pledges-net-neutrality-laws-if-elected-president/>

Cameron, G. (2013, December 12). U.S. appeals court strikes down FCC net neutrality rules.

Reuters. Retrieved from <http://www.reuters.com/article/us-usa-court-netneutrality-idUSBREA0D11420140114>

The Communications Act of 1934. (2013, November 27). Justice Information Sharing. Retrieved from <https://it.ojp.gov/PrivacyLiberty/authorities/statutes/1288>

Hu, E. (2014, September 14). 3.7 Million Comments Later, Here's Where Net Neutrality Stands.

NPR. Retrieved from

<http://www.npr.org/sections/alltechconsidered/2014/09/17/349243335/3-7-million-comments-later-heres-where-net-neutrality-stands>

Investor's Business Daily. (2014, November 11). Obama Pushes Net Neutrality For Control Of

Web. Retrieved from <http://www.investors.com/obama-pushes-for-internet-net-neutrality/>

Kang, C. (2011, January 20). Verizon sues FCC over net-neutrality rules. The Washington Post. Retrieved from <http://www.washingtonpost.com/wpdyn/content/article/2011/01/20/AR2011012005853.html>

CNN Politics. Retrieved from <http://www.cnn.com/2014/11/11/politics/fcc-chairman-protesters/>

Mechaber, E. (2014, November 10). President Obama Urges FCC to Implement Stronger Net

Neutrality Rules. [Web Log Comment]. Retrieved from <https://www.whitehouse.gov/blog/2014/11/10/president-obama-urges-fcc-implement-stronger-net-neutrality-rules>

Net Neutrality President Obama's Plan for a Free and Open Internet. (n.d.) The White House. Retrieved from <https://www.whitehouse.gov/net-neutrality>

The White House. (2014, November, 10). President Obama's Statement on Keeping the Internet Open and Free [Video file]. Retrieved from <https://www.youtube.com/watch?v=uKcjQPVwfDk>