The Many Dimensions of Net Neutrality
(and how we reduced them)

Erin Antono, Deger Turan, Justine Zhang

Introduction

In the spirit of democracy, government organizations often solicit public comments before a major ruling, ostensibly to incorporate the public's input into their final decision. The massive scale of comments these campaigns attract often prohibits manual inspection. The ability to analyze and understand this large body of comments computationally therefore allows these organizations to better understand the public's ideas and overall sentiment.

In this paper, we use unsupervised learning techniques to understand the content of the comments that the Federal Communications Commission (FCC) has received, relating to its upcoming decisions on net neutrality. In particular, we discuss the application of Singular Value Decomposition (SVD) to extract key topical and linguistic characteristics of comments. We first use SVD to extract common topics and their representative comments and keywords, based on a tf-idf weighting of each word used. We then use SVD on a smaller set of linguistic features, and discover five types of comments corresponding to the five top components extracted by SVD, which appear to reflect characteristics such as a commenter's personal investment in the issue and the other figures and organizations present in the comment's narrative. Finally, we perform a cross-comparison of the topical and linguistic components.

Data Set

As of September 2014, the FCC has received over 1.1 million comments on the subject of net neutrality, and has made the full set of comments available to the public. Before the start of our project, the Sunlight Foundation had done analysis on the comments, and made a set of cleaned data available to the public. For our project, we started with the Sunlight Foundation's set of 800,000 comments, which had been stripped of a large number of garbage comments[1]. To make the task more manageable, we constrained our analysis to a random subset of about 80,000 comments.

As was expected for a forum with open public submissions, the net neutrality comments contained large quantities of form letters submitted from a variety of sources. The Sunlight Foundation estimates that approximately 60% of the data consisted of form letters. This redundancy made analysis more difficult, because the large number of nearly-identical comments obscured much of the vocabulary and linguistic features present in the comments which did not originate from form letter campaigns. Hence we made an effort to remove these duplicates.

Preprocessing

We used a corpus of the original form letter documents from the Sunlight Foundation[1] and the nltk sentence tokenizer[2] to clean our data. We hence detected form letter content within each comment on a per-sentence basis, in order to retain original content which a user had added to a form letter. Finally, the sentences were normalized by removing punctuation and casing.

After early attempts at topic modeling, we found that our topic keywords were dominated by names and addresses; additionally, the most salient keywords of several topics all clearly originated from specific sentences in the comments which occurred frequently, and were missed by our intial form letter detection. To account for this problem, sentences containing irrelevant information such as names, addresses, or greetings were removed altogether. We further cleaned the data set by treating sentences that occurred more than 10 times as sentences from undetected form letters.

After this filtering, our results were still similarly impacted by the presence of short comments which were not necessarily from form letters, but were still nearly-identical to one another as a consequence of their length. While it may be interesting to study the propagation of such buzzwords as ”net neutrality” and ”level playing field” in short comments, we decided to focus our analysis on comments for which the writer could have plausibly contributed a meaningful amount of personal input, as opposed to terse comments which were basically variants of these buzzwords. As a simple heuristic, we removed all comments with less than four sentences. After these steps, our filtered dataset consisted of around 15,000 comments; among them there are 13 form letter comments.

Method

We used the Singular Value Decomposition algorithm (SVD) from scikit-learn [5] to perform dimensionality reduction on our dataset, allowing us to discover the most salient characteristics of the comments. As a brief overview of the process, given m comments and n features, let $X \in \mathbb{R}^{m \times n}$ be a matrix with $X_{i,j}$ = the value of feature j in comment i. Given some $k < n$, we produce a low-rank approximation of X,

$X_k = U_k \Sigma_k V_k^T$ where Σ_k is a diagonal matrix containing X's k largest singular values; this hence represents k components inferred from the original features. $U_{i,j}$ then corresponds to the value of new component j at comment i, while $V_{i,j}$ corresponds to the weighting of feature j at old feature i. Higher values roughly correspond to that component being more characteristic of the comment or feature.

Topic Analysis

To determine the most common topics represented in the comments, we used the term frequency-inverse document frequency (tf-idf) scores of each word in a particular comment as the set of features for that comment.

We applied the SVD algorithm to this set of features, which is also known as Latent Semantic Analysis (LSA) when used in this context.

Results

We chose the number of topics for interpretability of results, arriving at $k = 10$ topics (we also attempted to choose k such that the improvement in reconstruction error $|X - U_k \Sigma_k V_k^T|$ levelled off, producing 25 topics, but the differences between topics were much more subtle).

Using the notation above, after performing SVD, for each topic j, the most important words i maximize V_{ij} and the most representative comments maximize U_{ij}. For instance, here are the most representative words, and most representative comment for a selection of two topics:

Topic 2 - Cable Companies

Most important words: companies cable comcast speeds pay consumers charge speed competition netflix monopolies monopoly company money dont net big warner faster lane

Example comment: *"Regarding the proposal to permit cable companies and other ISPs to charge companies to provide faster downloads: This is the most ridiculous proposal I have heard in a long time. The cable companies, eg. Comcast and Time Warner are claiming that such fees to companies like Netflix would not create a two-tier fast lane-slow lane internet."*

Topic 3 - Legal terminology

Most important words: net neutrality fcc common title rules broadband ii carriers telecommunications protect reclassify proposed urge support wheeler chairman isps carrier public

Example comment: *"Net neutrality is the First Amendment of the Internet, the principle that Internet service providers (ISPs) treat all data equally. As an Internet user, net neutrality is vitally important to me. The FCC should use its Title II authority to protect it."*

Topic 6 - Small Business Equality

Most important words: fast lane lanes small business businesses slow big open internet pay playing field afford level rules innovation create equal traffic

Example comment: *"there is no difference between the statements 'fast lane' and 'slow lane', and 'fast lane' and 'hyper fast lane'. We're not stupid. This will only allow big corporations with money to squeeze out little start up companies that would have no way to afford to pay for this hyper fast lane..."*

The full list of topics we labelled after analyzing the examples are as follows: free and open Internet, America government and freedom, cable companies, legal terminology, innovation and startup encouragement, ISP profits, small business equality, ISP-consumer relationship, ISP-data treatment, companies controlling information.

Form letter topics

Given the high volume of form letters, one question we may ask is this: Do form letters talk about different topics than original comments, which are unaffiliated with any form letter campaigns? To answer this question, we use the topic weights for each comment as produced above. Consider the average topic weight of a set of comments S for topic j, $|S|^{-1} \sum_{x \in S} U_{xj}$. For each topic, we calculate the average weight over all the comments, over only the 13 form letter comments, and over a random sample of 13 comments (as a control). A bar graph of topic weights per topic is shown in Figure 1.

From the graph we can see that in particular, form letters seem to mention topic 3, "legal terminology", a lot more than the average comment (representative words and comments for this topic are listed above). This is corroborated by a manual examination of the form letter comments, which are mostly fairly explicit references to past legislation and proposed legislative changes:

"title ii of the communications act of 1934 already grants you the authority to declare the internet a public utility" (from the Daily Kos) [3]

"The FCC should use its Title II authority to protect [net neutrality]" (from Battle for the Net) [4]

Given these results, we may tentatively conclude the following: while most comments encompass a wider range of ideas, form letters are specifically organized calls for legislative change.

Linguistic analysis

In addition to determining the topic content of the comments, we also wanted to gain insight into how the comments were written. We hoped to capture differences in degree of personal investment, determine the intended audience and subject for different comments, and intensity and politeness of the commenters.

Features

We used features were based on linguistic characteristics of the comments, attitudes of the writers, and entities the comments were directed to. We considered the following features from each document:

% sentences with 1st singular related pronouns
% sentences with 1st plural related pronouns
% sentences with 2nd singular related pronoun
average # words per sentence
# sentences
% sentences with negation ("no", "nt", "not")
% sentences with "must" etc.
% sentences with "should" etc.
% sentences with "will" etc.
% sentences with "may" etc.
% sentences with "can" etc.
Flesch-Kincaid Reading Ease scores
Mentioning of Chairman Tom Wheeler
Mentioning of swear words
Mentioning of companies "ATT Warner Verizon Comcast"

The feature vectors, corresponding to rows of X from the above notation, were normalized to get a mean of 0 and variance of 1.

Results

Using these features, with $k = 5$, we found the following of comments which were characterized in surprisingly striking and intuitive ways, which we list below. The full table of components to features is shown in Figure 2.

1- Personal Worries - Defining Features: High use of I and negatives, low terminology, no directed audience:

"i was a computer wizard practically before i could read without a parent over my shoulder and i think its obvious why someone who knows computers so well would be so concerned..."

"dear fcc i use my pc like millions of others do online for research for diseases or for radio astronomy in conjunction with volunteer run projects at universities across the usa those that use the boinc interface, seti@home, einstein@home, rosetta@home, milkyway@home gpugrid etc"

2- Legal References - Defining Features: Low reading ease, low profanity, lengthy comments:

"forbearance furthers the objective of interpreting law in light of modern technology and markets without undermining its core purposes."

"before the federal communications commission washington dc in the matter of protecting and promoting the open internet gn docket no. 1428 framework for broadband internet service gn docket no. 10127 comments of comcast corporation comcast corporation"

3- Frustrated at Tom Wheeler - Defining Features: High use of you and profanity, directed at Tom Wheeler, concise comments:

"mr wheeler i will first remind you that you are an employee of the federal government of the united states of america. basically you work for us the people."

"mr wheeler as a paying customer of the internet i find what you are doing offensive and incredibly criminal. if you do not back down from this position of destroying net neutrality i and every one of the computer geeks i know will demand your resignation."

4-Dreams and Values - Defining Features: High use of plural pronouns, can, must; no directed audience

"please consider as this part of our era is economically hard in so many ways. for many like me who are somewhat housebound the internet is our library our bank"

"the internet has already greatly changed the way our world works and for a time this was acceptable. unfortunately legislation has failed to keep up with the technology and we have reached a crossroads that could make or break the continued prosperity and innovation the internet provides."

5-Experiences and Anecdotes -s Defining Features: High use of I, low negativity, short comments

"the internet is important to me because as someone who suffers from disabilities due to multiple sclerosis it gives me back some of my independence that this disease has stripped from me. i can go online and research treatments to better make informed decisions."

"i work for a company that creates comedic videos and puts them online. its my livelihood. if the internet becomes an exclusive club myself and many others may be out of a job."

When we used higher numbers of components, later components corresponding to small eigenvalues had a neutral and balanced amount of features, and were not very helpful in discerning commenter attitudes.

Overall Results

With these two different sets of characteristic components of our comment set, we wanted to investigate the relationship between the two. To do this, we created a matrix $M \in \mathbb{R}^{txc}$. t is the number of topics, c is the number of components, U_t is the U vector resulting from topic modeling and U_c is the U vector resulting from comment clustering.

$$M[i,j] = \sum_{k=0}^{n} U_t[k,i] * U_c[k,j].$$

From M, we obtain an other vector M', which is normalized such that $M' = M/max(M)$. A large $M'[i,j]$

value indicates high overlap between topic i and component j. A heatmap of these weightings is shown in Figure 3. We found that each component-vector pairing, which we call comment buckets, matched a characteristic portrait of a commenter.

The correspondence of the topics and components was reasonable. The most crowded comments buckets were:

1 - Personal worries - Innovation and startup encouragement. Most people in this bucket either are, or particularly concerned about small business and startup owners: *"i am a small businessman. the internet is critical my success. consigning me to a slow lane of the internet might do serious damage to the success of my business. also as a private individual i believe the internet serves as a public good"*

2 - Frustrated at Tom Wheeler - Equality for small business. The second most crowded bucket contains most of the comments with swearwords: *"if you dumb asses pass this law taxes will have to go up to pay for schools and because some schools wont be able to afford it so the education levels of schools will decrease... and facebook will die out and many stocks will drop and die affecting the stock market for many stockholders. with that all aside you will piss off millions for literally no reason"*

3 - Frustrated at Tom Wheeler - Government, America, Freedom. This bucket was much less profane, and used patriotic references and national values extensively: *"mr wheeler and fcc members i grow increasingly concerned with your attempts to ram an anticonsumer net neutrality bill through the process. we know what you are doing. you may feel inclined to bow to the corporate influences"*

4 - Legal References - ISPs Data Treatment. Containing relatively higher levels of rigor, this bucket shows that people who are worried about monopolization of data had the most eloquent comments: *"isps need to be reclassified as title ii common carriers allowing this proceeding to go through would allow isps to charge people extra fees to carry traffic from any online business that they want if a company depends heavily or entirely on internet traffic the isp could refuse to allow web pages to load in under a minute"*

5 - Experiences and Anecdotes - Innovation and startup encouragement. This bucket held an interestingly high number of comments from people who were very worried about their profits or lifestyles, but did not have a very clear understanding of the case: *"i am an artist. how can i succeed in an internet that favors already built giants. who will be able to find me in a segregated cyberspace. i will no longer be able to find endless inspiration and utilize the internet the way i do now"*

One bucket that we were surprised was the low correlation of use of legal terminology and legal references component. This may be the case because the legal references bucket had the length and rigor of the comment as a strong characteristic, and personal worries captured a lot of comments that were well informed and used terminology, but were short or included colloquialisms.

We found significant correlation between the legal reference comments, and the dreams and values comments in terms of subjects they tackled, even though their level of rigor is was distinctive. Large companies controlling information was repeatedly mentioned in personal worries, even though it is rarely mentioned by legal reference documents. Comments with high profanity were also simplest. Government and American values were rarely referred to in personal worries or experiences, but were very common on comments at Tom Wheeler, legal references and dreams and values.

Most of the garbage data that was unusually long, literary or complicated, such as the full text of The Great Gatsby or LCD screen instructions, were also contained in the legal references cluster.

Future Work

There are several directions in which we could expand our work from this paper. Looking further into our data set, we should be able to extract more information on how modifications to form letters correspond to the topics and comment characteristics that we have found. In our current dataset, which contains a fraction of the original comments, there were not enough form letters with original content added on to draw any conclusions.

We could also investigate the mapping additional features such as gender and location of the commenter or time of comment to the topics and characterizations we've identified. Triggers of high numbers of similar comment submissions can be deduced, which can be both in forms of celebrity or website encouragement, or response to events. It would be interesting to further study the application of similar analysis on a different body of comments, especially on subjects with multiple aspects and viewpoints. Since the data was 99% supporting net neutrality, the conclusions did not lead to a meaningful potential classification problem. The topic-component cross analysis can especially be helpful in debates with non-binary outcomes.

Conclusion

This research project is a first attempt at understanding the large and diverse data set that results from an open comment platform. By clustering similar attitudes, subjects and concerns, common threads can be found throughout the comments. Using this type of analysis, the voice of public sentiment can be better understood and taken into consideration for the issue of net neutrality, and other debates in the future.

Figures

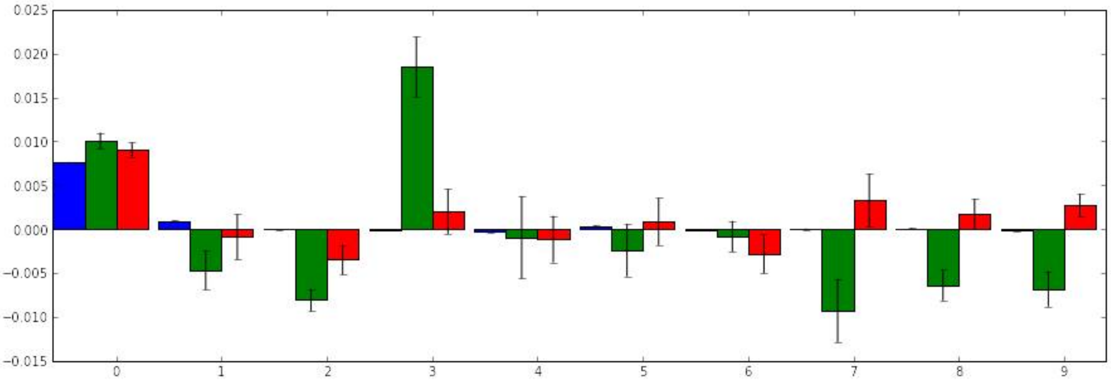

Figure 1: Topic Weights per Topic. For each of the three groups with blue corresponding to all comments, green corresponding to form letters, and red corresponding to random comments, along with standard error.

Feature	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5
I	0.343	0.082	0.141	-0.233	0.438
We	0.055	-0.154	-0.014	0.563	-0.381
You	0.008	0.353	0.564	-0.022	-0.153
Ave Len	0.619	-0.022	-0.041	0.029	-0.190
Tot Len	0.033	0.508	-0.423	0.195	0.086
Negative	0.452	-0.080	-0.097	-0.199	-0.166
Must	-0.067	-0.071	-0.121	0.271	-0.363
Should	0.012	-0.157	-0.354	-0.513	-0.250
Will	0.375	-0.028	0.139	0.170	0.039
May	0.218	0.016	0.017	0.096	0.064
Can	0.298	-0.034	0.046	0.201	0.081
Tom W	0.006	0.377	0.233	-0.150	-0.398
Cable	0.046	0.204	-0.120	-0.302	-0.423
Read Ease	-0.059	-0.508	0.419	-0.146	-0.153
Swear	-0.052	-0.328	0.256	-0.023	-0.016
Means:	0.153	0.054	0.043	-0.004	-0.119
Vars:	0.043	0.064	0.065	0.067	0.052

Figure 2: This table lists the weighing of different features for each component. The five components correspond to: Personal Worries, Legal References, Frustrated at Tom Wheeler, Dreams and Values, Experiences and Anecdotes.

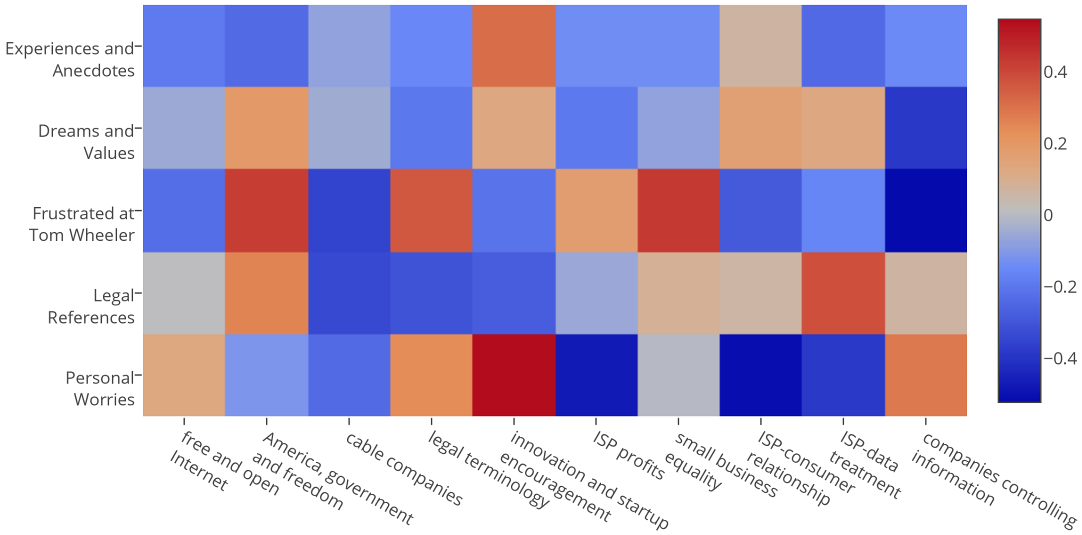

Figure 3: Heatmap of Topics and Components. This figure was generated using plot.ly[6]

References

1. Lannon, Bob, and Andrew Pendleton. "What Can We Learn from 800,000 Public Comments on the FCC's Net Neutrality Plan?" Sunlight Foundation, 2 Sept. 2014. Web. 08 Dec. 2014. http://sunlightfoundation.com/blog/2014/09/02/what-can-we-learn-from-800000-public-comments-on-the-fccs-net-neutrality-plan/

2. Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. OReilly Media Inc.

3. "Sign the Petition: Save Net Neutrality, Stop the Dangerous." Daily Kos. Web. 13 Dec. 2014. https://www.dailykos.com/campaigns/785

4. "This Is Why Your Internet Is Slow. And It'll Get Worse. Unless You Take 1 Min to Do This, Now." Battle For The Net. Team@fightforthefuture.org. Web. 13 Dec. 2014. https://www.battleforthenet.com/ 5. Pedregosa et al., Scikit-learn: Machine Learning in Python,, JMLR 12, pp. 2825-2830, 2011. http://scikit-learn.org/stable/modules/decomposition.html

6. "Heatmap Made by Eantono — Plotly." Heatmap Made by Eantono @ Plotly. Plot.ly at https://github.com/plotly, 13 Dec. 2014. Web. 13 Dec. 2014. https://plot.ly/ eantono/28