Topic Analysis of the FCC's Public Comments on Net Neutrality

Sachin Padmanabhan spadman@stanford.edu
Leon Yao leonyao@stanford.edu
Luda Zhao ludazhao@stanford.edu
Timothy Lee timothyl@stanford.edu

Department of Computer Science
Stanford University

Abstract

The FCC's proposed net neutrality policy change in 2014 was met with widespread public controversy and outrage. The FCC recently released to the public millions of comments that it received about the issue. It is abundantly clear that the vast majority of citizens prefer to have net neutrality intact, but what exactly are the people saying? What are their main arguments and reasons for wanting to maintain net neutrality? In this project, we use natural language processing techniques to analyze the arguments in 800,000 of the comments.

1. Introduction

1.1. Motivation

The Open Internet proceeding of the FCC (Federal Communication Commissions) is a critical regulatory effort to determine the future of the Internet. The proceeding concerns "Net Neutrality", the principle that all Internet traffic should be treated equally, and no internet provider will be given control over Internet traffic. Should the FCC decide not to maintain net neutrality, ISPs will have more power to regulate Internet traffic and scrutinize data sent over the Internet. In addition, ISPs will be able to discriminate between Internet traffic to provide a "fast lane" for high-paying consumers. Proponents of net neutrality argue that this will severely restrict free speech and privacy on the Internet. In addition, they assert that giving ISPs differential control over Internet traffic will ultimately result in extremely slow Internet for average consumers, including individuals and corporations, that cannot afford to pay as much as large corporations. In turn, this will hamper fair competition between businesses, shifting the balance largely in the side of large companies to the immense detriment of innovative startups. Instead, proponents maintain that the Internet should instead be reclassified as a "common carrier".

Project done in Stanford University's CS 229 (Machine Learning) course taught in Autumn 2014 by Professor Andrew Ng.

The debate on net neutrality has attracted a large response from the American public. As of writing, the FCC proceedings has attracted over 2 million comments. These comments are important to the FCC's decision making process, and are usually read by people. However, given the unprecedented number of comments, this is not possible. Under this context, natural language processing is an effective technique that can help gain insight into the comments as a whole. Can we automatically determine which issues were most pertinent to proponents of net neutrality?

1.2. Our Work

After reading many of the comments, we saw that the arguments were almost unanimously in favor of maintaining net neutrality. However, the arguments presented varied greatly in length, relevance, level of insight, and topic. We wanted to determine what people's arguments are and *why* they are in favor of net neutrality. The majority of the comments made at least one of the following arguments:

- Net neutrality is needed to protect freedom of ideas, creativity, speech, and communication on the Internet (`ideafreedom`)
- Net neutrality is needed to protect fair market competition for small businesses and startups (`fairbusiness`)
- Net neutrality is needed to protect the Internet from further legislation and government intervention in the future (`fairgov`)

Our goal was to classify the argument of each comment into one or more of the above topics using supervised learning.

We decided against using an unsupervised learning approach since it was attempted earlier by a team at Sunlight Labs. The results they got were not very inspiring because the clusters they found were only indicated by a few key words, but were too vague to have any concrete topic behind it. Among some very bad clusters, we did see some interesting ones like, "small market," "bidding," "premium," and "disadvantage." This exactly fits our idea of free business. Instead of having 1 in nearly 100 clusters be interesting, we thought it would

be interesting to remove the noise and just focus on 3 topics we knew were in the dataset.

2. Data

The original dataset released by the FCC consisted of 1.1 million raw comments along with metadata, many of which were blank, unparseable, or too long (*Les Misérables* and *War and Peace* were both submitted as comments). Fortunately, the team at Sunlight Labs processed the dataset to remove these unworkable comments and provided a cleaner dataset of 800,959 comments with metadata in JSON format.

```
{
  "applicant": "Kara J. Walton",
  "dateRcpt": "2014-06-16T04:00:00Z",
  "stateCd": "VA",
  "zip": "20121"
  "text": "Allowing the cable companies to
           start charging companies for..."
}
```

To train and test our classifier, we drew a random sample of 800 comments from the dataset. We manually read through each comment to discern the arguments presented and correspondingly labeled each comment.

```
{
  ...
  "topiclabels": {
    "ideafreedom": 1,
    "fairbusiness": 1,
    "freegov": 0
  },
  "formletter": 0,
  "personal": 0
}
```

The rest of the comments were used after we built the classifier to glean interesting insight on the entire dataset.

3. Methodology

3.1. Form Letter Detection

By reading through the dataset, we noticed that a majority (about 60%) of it was composed of "form letters," which were mostly identical comments written by third party organization who had its supporters send in the same professionally written messages.

> To Chairman Tom Wheeler and the FCC Commissioners To the FCC Please build any net neutrality argument upon solid legal standing. Specifically, this means reclassifying broadband under Title II of the Telecommunications Act of 1934. 706 authority from the Telecommunications Act has been repeatedly struck down in court after legal challenges by telecom companies. Take the appropriate steps to prevent this from happening again. Sincerely, XXXX

Clearly, a form letter could often times sound exactly like a regular comment. We decided to use form letters in our topic classification because, despite their spam-like nature, they still signify the intentions of the individual sender who agrees with this mass message, otherwise they wouldn't have taken the time to actually send it. We found that most, if not nearly all, of these messages were from different people. Thus, one of the main problems we had to tune for was overfitting the training set.

Instead, our goal was to use unsupervised learning to detect exactly which comments were form letters so that we could perform analysis on just the form letters themselves. We did this using the Simhash algorithm, which is a generally fast method to calculate the similarity between two documents, and is effective for near-duplicate detection.

Using the Simhash algorithm, we we found the near-duplicates to the document being classified. If there existed a significant number of comments that were near duplicates, then the comment was classified as a form letter. We used a 64-bit hash size, shingle width of 4 letters, and hamming distance threshold of 10 bits as parameters for the model. Given a labelled data set of 800 comments, the model classified form letters with 88% accuracy. On a data set containing 40,000 comments, the model showed that 63% of the comments were form letters, similar to our initial observation's 60% proportion.

3.2. Feature Selection

We first preprocessed the data by removing stop words such as "the," "a," "and," etc. that appear in nearly all comments but are essentially useless features. We also stemmed our words, so that different conjugations of the same word would be counted as the same. We also tried using different size n-grams to increase our feature space and to capture more of the word contexts.

We used several standard features for typical NLP datasets. We first found the word counts of our comments, then normalized them and used TF–IDF (Term Frequency–Inverse Document Frequency) features, which is a weighting factor for each word that gives a value proportional to the frequency of that word in the comment offset by the frequency of the word in the entire corpus. This allowed us to remove words that appear in every comment, but are bad features to use for training a classifier. For example, words like "Internet" and "FCC" were used in nearly every comment, but are not helpful for determining if a comment is from a given class. TF–IDF allows us to hone in on the most important features,

which is one of the best methods for feature selection.

$$\mathrm{tf}(t,d) = 0.5 + \frac{0.5 f(t,d)}{\max\{f(w,d) : w \in d\}}$$
$$\mathrm{idf}(t,D) = \log \frac{N}{|\{d \in D : t \in D\}|}$$
$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t,D)$$

On top of TF–IDF we also used a min/max frequency pruning. If a word only appears once or twice in the dataset or in every single comment, TF–IDF will assign it a low score, but we wanted to actually reduce our feature space so as to not overfit. If a word has word frequency less than our min or greater than our max, then we removed it.

3.3. Model Selection

For each of our classifiers we learned a one vs. all classifier because a particular comment could have multiple different topics. We used 10-fold cross validation for each model, so each training/testing accuracy we report are the generalization accuracies.

3.3.1. Bernoulli Naïve Bayes Classifier

The first classifier we tried was just a simple Naïve Bayes with Laplace smoothing for data distributed according to the Bernoulli distribution. By finding the maximum likelihood estimates

$$\phi_{j|y=1} = \frac{\sum_{i=1}^{m} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 1\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}$$
$$\phi_{j|y=0} = \frac{\sum_{i=1}^{m} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 0\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}}$$
$$\phi_y = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}{m}$$

and then determining the class with the highest posterior probability, we obtained the results in Table 1. We saw that

Table 1. Classification accuracies for Naïve Bayes classifier

Topic	Training	Testing
ideafreedom	87.14%	85.70%
fairbusiness	86.60%	80.76%
freegov	96.19%	96.46%

the Naïve Bayes Classifier suffered a lot from the form letters and also overfitted the training set.

3.3.2. Regularized (Bayesian) Logistic Regression

Since overfitting was a problem for Naïve Bayes, we decided to use regularization to restrict the norm of the learned parameters to control the VC dimension of our classifier. We used logistic regression with ℓ_2 regularization, which corresponds to a Gaussian prior on the data. Thus, we implemented a stochastic gradient descent classifier to minimize the cost function

$$\theta = \arg\max_{\theta} J(\theta) = \sum_{i=1}^{m} \log p(y^{(i)}|x^{(i)};\theta) - \frac{\lambda}{2}\|\theta\|_2^2$$

The results are summarized in Table 2.

Table 2. Classification accuracies for ell_2-regularized logistic regression

Topic	Training	Testing
ideafreedom	99.24%	90.89%
fairbusiness	99.59%	87.72%
freegov	99.24%	96.46%

3.3.3. Support Vector Machine

We finally tried an ℓ_1-norm soft margin SVM classifier with a Gaussian kernel.

$$\begin{aligned} \min_{\alpha} \quad & \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j K(x^{(i)}, x^{(j)}) \\ \text{s.t.} \quad & 0 \le \alpha_i \le C,\ i = 1, \ldots, m \\ & \sum_{i=1}^{m} \alpha_i y^{(i)} = 0 \end{aligned}$$

$$K(x,z) = \exp\left(\frac{\|x-z\|^2}{2\tau}\right)$$

Although computationally more intensive, we felt it would yield better results. Indeed, chosen with default parameters, this gave better results than the previous methods. In order to further improve the results, we ran a model selection algorithm to search for the best parameters for the model, and the resulting classifier yielded even better results. The results for the optimized classifier are summarized in Table 3.

Table 3. Classification accuracies for support vector machine with Gaussian kernel

Topic	Training	Testing
ideafreedom	95.22%	90.92%
fairbusiness	97.37%	89.03%
freegov	97.45%	97.48%

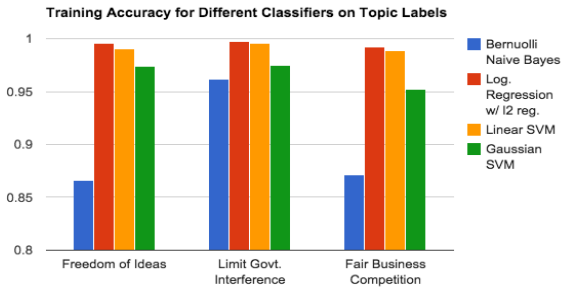

Figure 1. Training Accuracy

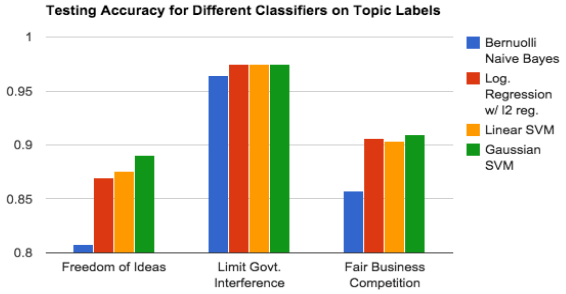

Figure 2. Testing Accuracy

3.4. Evaluation

After selecting our best classifier, an SVM with Gaussian kernel, we also looked at the precision and recall statistics in terms of a confusion matrix, where each column of the matrix represents the instances in a predicted class (negative, positive), while each row represents the instances in an actual class (negative, positive). Splitting data half and half for training and testing, we obtained the following confusion matrix for each topic:

ideafreedom

$$\begin{bmatrix} 168 & 10 \\ 39 & 179 \end{bmatrix}$$

Precision: 81.2%
Recall: 94.4%

fairbusiness

$$\begin{bmatrix} 178 & 18 \\ 25 & 175 \end{bmatrix}$$

Precision: 90.8%
Recall: 87.7%

fairgov

$$\begin{bmatrix} 386 & 0 \\ 9 & 1 \end{bmatrix}$$

Precision: 100%
Recall: 97.7%

We see that for all of our topics, the classifier achieved both high precision and high recall. This result boosts our confidence that this particular classifier will be able to obtain a reasonable classification on our unlabeled data.

4. Results & Analysis

Since the SVM with a Gaussian kernel was our best classifier, we used this classifier to derive insights on the entire dataset of comments.

Overall, we found that the vast majority of comments talked about either the idea of freedom or fair business practices, with the plurality of these comments mentioning both arguments. These results are shown in the Venn diagram in Figure 3.

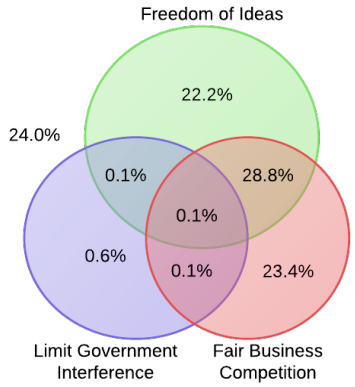

Figure 3. Distribution of topics among all comments

In addition, we thought it would be interesting to analyze the arguments that people in different states made. California had about 24% of the total comments, and they lead the percentage in every topic as well. So in order to see the the topic breakdown for specific states, we found the percentage of each state that argued about each topic. For example, only 0.06% of California talked about freedom from government interference while 0.5% of Florida and 0.41% of Texas talked about it. Although "free government" was not a very common topic talked about, Florida and Texas still had a significant number of people talk about it compared to the overall percentage of "free government" comments in the dataset (0.17%), considering about 84% of comments that argued about "free government" did not specify a state.

These results are interesting because Texas and Florida are both predominantly Republican states compared to California, so people in these states would care more about the traditionally Republican ideal of having a small government. This directly corresponds to their distaste towards government involvement with net neutrality.

Additionally, California accounts for about 24% of all the comments, but accounts for about 39% of the "idea freedom" topic. This indicates that a majority of California cares about topics relating to our first amendment rights and the ability to freely post things on the internet. This makes sense because California is one of the most liberal states in the country.

Other than these two anomalies, other states' topic distributions were mostly the same as the overall topic distribution.

5. Conclusion

Through our investigation, we've gained a better understanding of the issues that people have raised regarding net neu-

trality in the FCC's public comments. By identifying the most prominent concerns and training a classifier using a pre-labeled training set, we were able to classify 800,000 comments and capture their broad sentiments using a fraction of time and manpower as traditional procedures of review. In addition, we were able to apply our topic classification labels to make interesting observations about the geographical distribution of topics, in which we found out that distribution of our topic seems to follow certain regional political trends, an unexpected but fascinating result.

Furthermore, as with most publicly gathered comments, our dataset contains a large percentage of largely identical form letters. Since they provide a useful metric of the level of active public participation, we found it a worthwhile endeavor to identify them. We were fairly successful in this regard in using the Simhash algorithm, as our predicted percentages of 63% matched closely with the actual amount.

6. Future Work

From this project, there are many multiple directions we can take to abstract information from the comments for a deeper analysis. For example, the analysis of comments pertaining to form letters could provide very useful information. After finding the clusters of comments from form letters in the dataset, we can observe the geographic origins of form letters. In addition, we can apply the same model to perform topic classification on the set of comments of form letters only. Also, we can group comments by time to see what events cause form letters to be sent (for example, a television advertisement impels viewers to send the comment through a website). Besides form letters, we can look at the comments at a finer granularity through the lens gender. We can apply the same model to perform topic classification to sets of comments from different genders. By continuing this work, we hope to achieve more interesting results about the public's perception of net neutrality.

Acknowledgments

Special thanks to Professor Dan Jurafsky.

References

Andoni, Alexandr and Indyk, Piotr. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pp. 459–468. IEEE, 2006.

Bishop, Christopher M et al. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

Charikar, Moses S. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pp. 380–388. ACM, 2002.

Gong, Caichun, Huang, Yulan, Cheng, Xueqi, and Bai, Shuo. Detecting near-duplicates in large-scale short text databases. In *Advances in Knowledge Discovery and Data Mining*, pp. 877–883. Springer, 2008.

Hastie, Trevor, Tibshirani, Robert, Friedman, Jerome, Hastie, T, Friedman, J, and Tibshirani, R. *The elements of statistical learning*, volume 2. Springer, 2009.

Lannon, Bob. What can we learn from 800,000 public comments on the fcc's net neutrality plan? *Sunlight Foundation Blog*, 2014.

Manning, Christopher D, Raghavan, Prabhakar, and Schütze, Hinrich. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Russell, Stuart and Norvig, Peter. Artificial intelligence: A modern approach. *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs*, 25, 1995.