STAT 689 PROJECT REPORT

Instructor : James Long

Analyzing Public Comments to the FCC on Net Neutrality Using Agglomerative Clustering and Text Identifier Classification

Jacqueline Antwi-Danso[1,2]* | **Tarini Konchady**[1,2]*

[1]Department of Physics & Astronomy, Texas A& M University, 4242 TAMU, College Station, TX, 78743

[2]George P. and Cynthia Woods Mitchell Institute for Fundamental Physics and Astronomy, Texas A&M University, College Station, TX, 78743

Correspondence
Email: jadanso@tamu.edu
Email: tkonchady@tamu.edu

Funding information
N/A

Shortly after the Federal Communications Commission opened a public portal for the submission of comments concerning net neutrality, a number of studies and preliminary analyses reported evidence of spamming. Using agglomerative clustering and a binomial model, we analyzed a subset of these comments. We found that the success probability, p, combined with cluster silhouette scores, are adequate metrics for distinguishing between genuine and fake comments.

1 | INTRODUCTION

On December 14, 2017 the Federal Communications Commission (FCC) voted to repeal Title II oversight of Internet Service Providers (ISPs) [1]. This was a great blow to supporters of net neutrality, which is the idea "that Internet providers should be neutral gateways that provide equal access to all legal web content" [1]. Prior to the vote, the FCC had a docket for public comments on the vote open from April 26, 2017 till August 16, 2017 [2]. The FCC received nearly 24 million comments to the docket, which spoke to the public outcry.

1.1 | Motivation

A disturbingly large number of the public comments appeared to come from fraudulent sources: red flags pointing to this included comments having the same name, bulk simultaneous comment submission, and identical messages. The Pew Research Center reported that only 6% of the submitted comments were unique [3]. New York Attorney General Eric Schneiderman's office also conducted an investigation and determined that as many as 2 million comments were faked [4].

* Equally contributing authors.

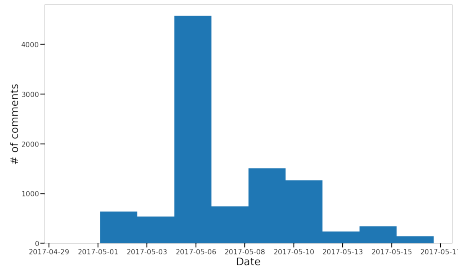


FIGURE 1 Comments submitted over time. This plot accounts for 10,000 comments, the entirety of our data.

A study that was of particular interest to us was Jeff Kao's analysis of over 22 million comments [5]. Kao used natural language processing techniques to tease out large swathes of pro-repeal comments. He presented three key findings: one, over a million comments appeared to use mail-merge to generate very similarly phrased pro-repeal comments; two, there were likely several pro-repeal campaigns that aimed to put possibly millions of duplicate submissions into the pool; and three, likely 99% of the unique comments were against the repeal. The code Kao used is available on GitHub.

2 | OBJECTIVE

We aimed to analyze comment text and identifiers, and attempt to quantify how likely it was that a given comment was fraudulent based on its content and identifiers. We worked almost exclusively in Python.

2.1 | Our Data

While cleaned datasets like the one Kao used were available, we chose to work with the FCC-provided data [6]. This is because the former were de-identified. The comments were divided into sets of 10,000 based on submission time (see Figure 1). The data provided 28 columns, and we found the following to be most relevant:

- Comment ID
- First and last name of filer
- Filer address
- Filer email address
- Time of submission
- Whether or not filer opted for email confirmation
- Comment text

3 | ANALYSIS

3.1 | Text Clustering

To classify the text of the comments, Kao vectorized then clustered the text. We used the same software to vectorize, namely spaCy. spaCy is a natural language processor, and is considered one of the fastest syntactic parsers in the world [7]. Like Kao, we used spaCy to vectorize our comments.

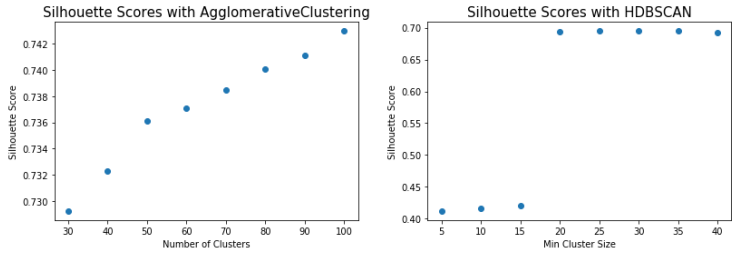


FIGURE 2 Silhouette scores from AgglomerativeClustering (left) and HDBSCAN (right).

Vectorizing text involves not only assigning a number to a word based on its value (say, giving "dog" a different number than "cat") but taking into account the proximity of the word to the words around it. That is to say, "the cat" would give a different value to "cat" than would "a cat".

After vectorizing the text, we clustered it. Kao suggested two clusterers — scikit-learn's AgglomerativeClustering [8], and hdbscan [9]. AgglomerativeClustering is an agglomerative clusterer as the name suggests. It starts by assuming each data point is in its own cluster and merges similar clusters till the resulting larger clusters are too dissimilar to merge. HDBSCAN is a hierarchical extension of density-based spatial clustering of applications with noise (DBSCAN). It examines the data and creates clusters based on how densely the data is clumped. To choose which clusterer to use, we first computed silhouette scores for the two.

A silhouette score denotes how similar or dissimilar a data point is to the rest of its cluster [10]. Silhouette scores can range from -1 to 1, with 1 indicating that the data point is very similar to the rest of its cluster and -1 indicating that it is very dissimilar to the rest of the cluster. To calculate silhouette scores, the following equation is used,

$$s_i = \frac{b(i) - a(i)}{max\{b(i), a(i)\}}$$

where s_i is the silhouette score of point i, $a(i)$ is the lowest average distance between point i and all the other points in its cluster, and $b(i)$ is the lowest average distance between point i and all the other points not in its cluster.

AgglomerativeClustering requires the user to set a number of clusters before it begins clustering the data while HDBSCAN requires a minimum cluster size. We considered a range of clusters and cluster sizes, and calculated the resulting silhouette scores (see Figure 2). We chose AgglomerativeClustering as it yielded the higher silhouette score. We chose our number of clusters to be 50 in order to avoid overfitting the comments.

After clustering, we calculated and plotted silhouette scores to visualize the cluster sizes (see Figure 3). To determine whether comments were pro- or anti-net neutrality we examined random comments within each cluster. These results are shown in Figure 4. The most standout feature is cluster 35, which was anti-net neutrality and contained 4,062 identical comments.

3.2 | Identifier Classification

The second aspect of our data analysis involved generating a metric for the "spamminess" of each comment by taking into consideration the identifying information. This is what differentiates our study from the others. A clustering-only approach to the classification of the comments gives a general, but not holistic overview of the source and stance of the comments. That is because the FCC allowed bulk submission methods via and API and a CSV template, which could

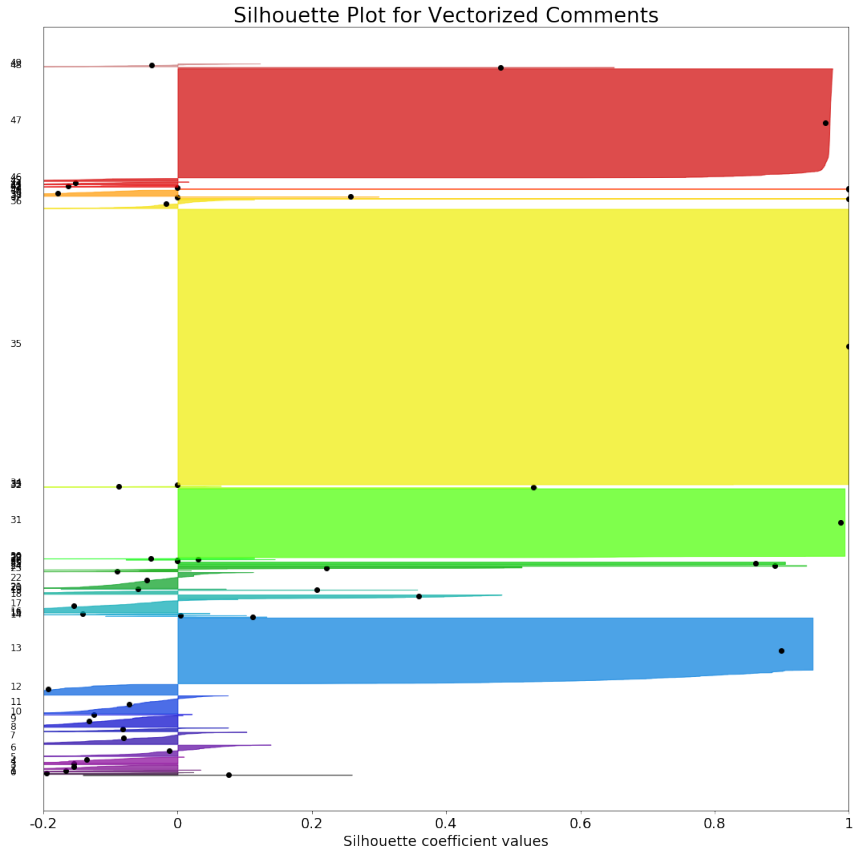


FIGURE 3 Silhouette scores from AgglomerativeClustering. The mean silhouette score of each cluster are marked by the black points, and the cluster IDs are on the y-axis. Several clusters appear to be very similar or very dissimilar. Only two of the clusters were anti-net neutrality.

be taken advantage of to submit spam. In addition, many advocacy groups created form letters which were used by their members in filing comments. We cross-checked the identifying information of each comment against the following sources:

- IMDB list of 100 most popular celebrities in the world [11]
- Breached accounts from HaveIBeenPwned [12]
- Form letters, as well as John Oliver's recommended comments [13, 14, 15, 16, 17, 18, 19, 20]
- Fight for the Future list of impersonated people [21]

3.2.1 | Natural Model

For each of the identifying columns, we generated a series of answers to yes/no questions:

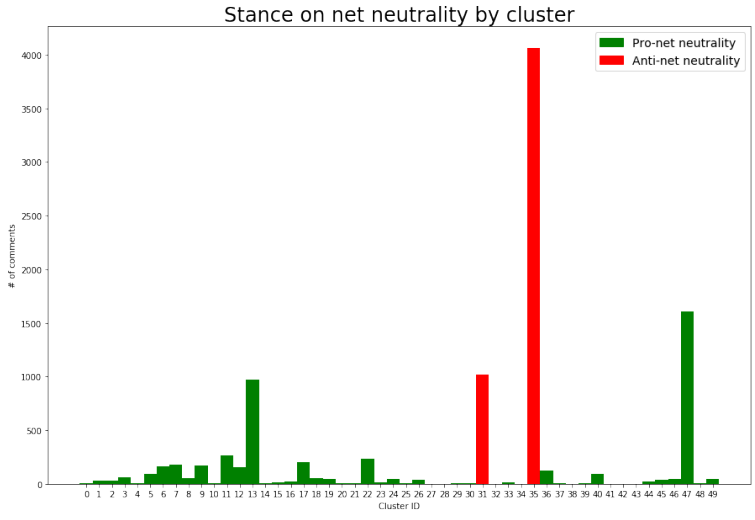


FIGURE 4 Stances of clusters on net neutrality.

- Is the first name realistic?
- Is the last name realistic?
- Do the name and email address match?
- Is this filer's name a celebrity name?
- Does the filer name and address information match that of an impersonated person?
- Is the comment from a form letter?
- Was the filer's email address involved in a breach?
- Did the filer opt for email confirmation?
- Did the filer supply complete address information?
- Do the filer's name and email address match those of any other filer(s)?

The natural mathematical model for this framework is a Bernoulli trial, such that for n number of questions x number of successes, and success probability p, we obtain a Binomial distribution, which has the following probability mass function:

$$P(x; p, n) = \left(\begin{array}{c} n \\ x \end{array} \right) (p)^x (1-p)^{(n-x)} \qquad \text{for } x = 0, 1, 2, \cdots, n$$

where

$$\left(\begin{array}{c} n \\ x \end{array} \right) = \frac{n!}{x!(n-x)!}$$

and

$$p = \frac{x}{n}$$

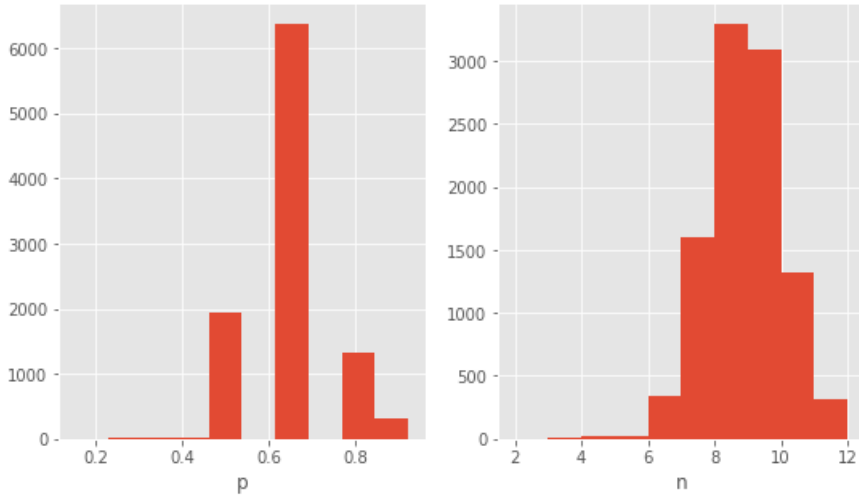


FIGURE 5 Distribution of success probability (left) and number of successes per question (right) for comments in our sample.

In order to make the model intuitive, such that fake comments always have low p and genuine ones have high p, we grouped the trial questions into positive and negative. For instance, a positive comment feature is having a name and email address that match and an examples of a negative one is having an email address that was involved in a breach. We assigned a score of 1 if a comment passed a positive question, and 0 if it passed a negative question.

Bayesian Analysis Informed by Kao's study, where it was found that approximately 10% all comments were non-spam and non-campaign (not from a form letter) we chose a beta prior with $\alpha = 10$ and $\beta = 90$.

$$f(x) = \frac{(x-a)^{p-1}(b-x)^{q-1}}{B(p,q)(b-a)^{p+q-1}} \qquad a \leq x \leq b; p, q > 0$$

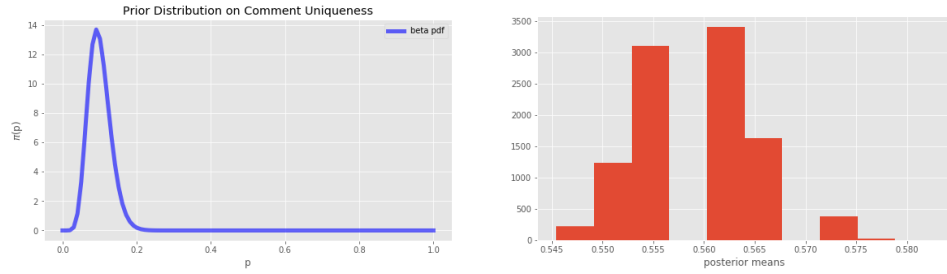
where

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$$

is the Beta function. We used the standard form of the prior ($a = 0$, $b = 1$ and $p = \alpha$, $q = \beta$). Our likelihood is given by the multiplicative sum of the Binomial distribution from above, and the posterior determined by conjugacy:

$$\alpha' = \alpha + \sum_{i=1}^{n} x_i$$

$$\beta' = \beta + n - \sum_{i=1}^{n} x_i$$



(a) (b)

FIGURE 6 Prior on comment uniqueness and distribution of posterior means of comments.

Using Baye's theorem, we calculated the posterior means, shown in Figure 6. The posterior means span a narrow range : ~ 0.54 - 0.58. This suggests that our prior may be too restrictive, or that we may need to choose another prior form. Furthermore, all of the comments have posterior means greater than 0.5, which is unrealistic. Due to this, we decided to use the success probability, p, as the metric for the probability that a comment is spam instead of the posterior means.

4 | RESULTS AND FUTURE WORK

We investigated whether or not there was a correlation between the cluster IDs we obtained from agglomerative clustering and success probability. We plot this in Figure 7. There is no correlation between p and cluster ID, which suggests that with p we are able to distinguish between spam and genuine comments that may have semantic similarity. This means that p can be used to distinguish between comments in the same cluster that have different qualities of identifying information.

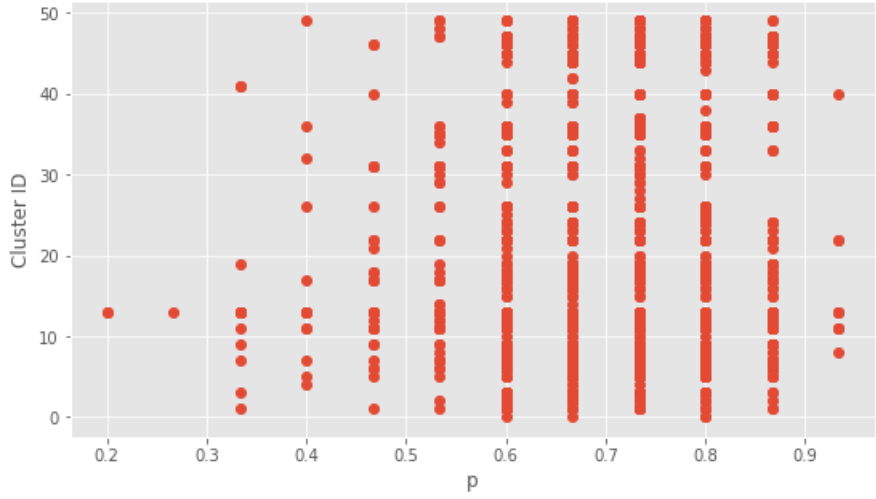


FIGURE 7 Comment cluster ID as a function of success probability, p.

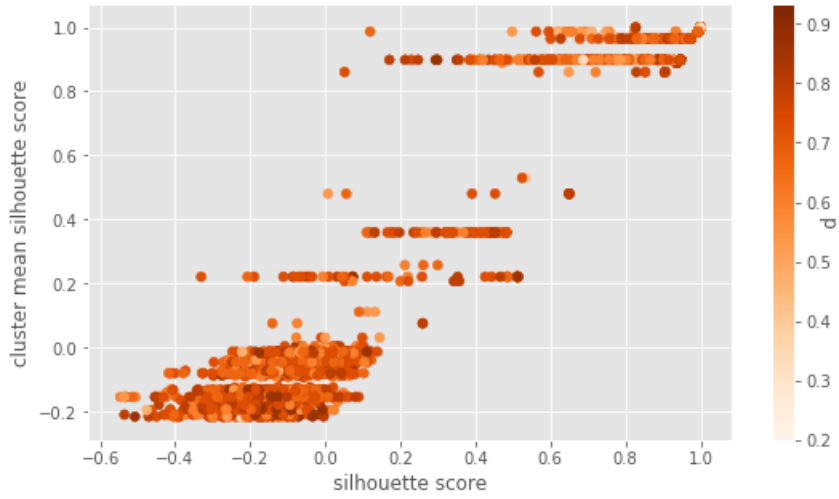


FIGURE 8 Cluster mean silhouette score as a function of individual comment silhouette score, color-coded by success probability

Figure 8 shows another way of visualizing the performance of p. Since the silhouette score is a measure of how similar a comment is to others in the same cluster based on the chosen distance metric, we see here again that there is a range of success probabilities for the different cluster assignments.

Figure 9 is the same as Figure 8, but color-coded according to stance on net-neutrality, which we determined using the cluster assignments. The pro-NN comments have low silhouette scores, which suggests that they have less semantic similarity to other comments in the sample, while anti-NN comments are very well clustered. This appears to corroborate the findings of other studies, suggesting that most of the campaigns came from anti net neutrality sources.

Now, we discuss the limitations of our model and ways to improve it. One issue is that it may be too simplistic. We could remedy this by generating a model that assigns a greater weight to certain test questions. Secondly, the model is inaccurate with little information. For example, a comment with some fields missing may have a similar success probability as one with more negative features than good, although the former may be genuine. We could remedy this by obtaining more external data to cross-check our identifying information with, such as census records. Next, spaCy's Named Entity Recognition and natural language processing features have an accuracy of 85%. However, by training it on another subset of the FCC dataset, we could greatly improve this. Lastly using a metric for determining the number of cluster assignments to use, such as the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) will help safeguard against overfitting.

5 | SUMMARY

We analyzed a sample of 10,000 comments on net neutrality from the Federal Communications Commission. Using natural language processing tool, spaCy and agglomerative clustering, we classified the comments into groups based on semantic similarity. We found that the Binomial distribution parameter, p, as an adequate metric for the quanifying the probability that a comment is either spam or genuine. In addition, we found a general correlation between net neutrality stance and the likelihood that a comment is from a campaign, which supports the findings from previous studies. Our

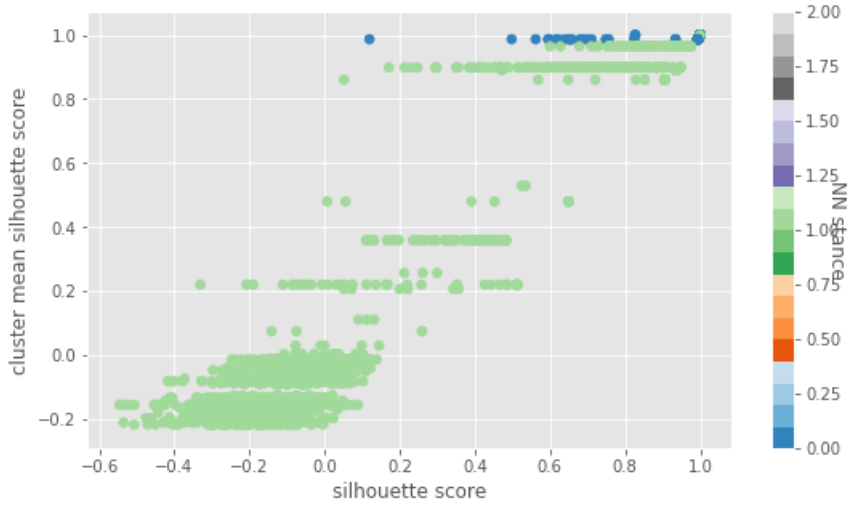


FIGURE 9 Cluster mean silhouette score as a function of individual comment silhouette score, color-coded by NN-stance. Green : Pro-NN; Blue : Anti-NN.

code is available on GitHub

REFERENCES

[1] Selyukh A, FCC Repeals 'Net Neutrality' Rules For Internet Providers; 2017. www.npr.org/sections/thetwo-way/2017/12/14/570526390/fcc-repeals-net-neutrality-rules-for-internet-providers.

[2] Brodkin J, How Netflix, reddit—and even Comcast—pledged support for net neutrality today; 2017. arstechnica.com/information-technology/2017/05/after-net-neutrality-comment-system-fails-senators-demand-answers/.

[3] Hittlin P OK, S T, Public Comments to the Federal Communications Commission About Net Neutrality Contain Many Inaccuracies and Duplicates; 2017. www.pewinternet.org/2017/11/29/public-comments-to-the-federal-communications-commission-about-net-neutrality-contain-many-inaccuracies-and-duplicates/.

[4] Naylor B, As FCC Prepares Net-Neutrality Vote, Study Finds Millions of Fake Comments; 2017. www.npr.org/2017/12/14/570262688/as-fcc-prepares-net-neutrality-vote-study-finds-millions-of-fake-comments.

[5] Kao J, More than a Million Pro-Repeal Net Neutrality Comments were Likely Faked; 2017. hackernoon.com/more-than-a-million-pro-repeal-net-neutrality-comments-were-likely-faked-e9f0e3ed36a6.

[6] Commission FC, FCC FACILITATES REVIEW OF RESTORING INTERNET FREEDOM RECORD; 2017. apps.fcc.gov/edocs_public/attachmatch/DA-17-1089A1.pdf.

[7] AI E, spaCy; 2018. https://spacy.io/.

[8] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011;12:2825–2830.

[9] McInnes L, Healy J, Astels S. hdbscan: Hierarchical density based clustering. The Journal of Open Source Software 2017 mar;2(11). https://doi.org/10.21105%2Fjoss.00205.

[10] Rousseeuw P. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 1987;.

[11] Bharathwiki, 100 MOST POPULAR CELEBRITIES IN THE WORLD; 2013. https://www.imdb.com/list/ls052283250/.

[12] HaveIBeenPwned.com;. https://haveibeenpwned.com/.

[13] ;. http://act.demandprogress.org/letter/net_neutrality_sen/.

[14] ;. https://act.openmedia.org/NetNeutralityDearEditor#newmode-embed-2-1313.

[15] Battle For The Net;. https://www.battleforthenet.com/letter/.

[16] Trayser L, 7 Net Neutrality Letters You Can Send to Resistbot today; 2017. https://medium.com/words-for-life/7-net-neutrality-letters-you-can-send-to-resistbot-today-7ff46d772afe.

[17] Foundation EF;. https://act.eff.org/action/save-the-open-internet-orderfreepress.net:https://act.freepress.net/letter/two_million/?

[18] Staff C, CFIF Mobilizes Americans Opposed to the Obama Administration's Title II Internet Power Grab; 2017. http://cfif.org/v/index.php/commentary/62-technology-and-telecom/3596-center-for-individual-freedom-mobilizes-americans-opposed-to-the-obama-administrations-title-ii-internet-power-grab-.

[19] ;. https://www.protectingtaxpayers.org/take-action/.

[20] ;. https://action.americancommitment.org/ctas/advocacy-251-repeal-obamas-internet-regulations/letter?zip=11249.

[21] Letter to the FCC from people whose names and addresses were used to submit fake comments against net neutrality; 2017. https://www.fightforthefuture.org/news/2017-05-25-letter-to-the-fcc-from-people-whose-names-and/.