# Analysis of FCC Net Neutrality Submissions

Fake or Not using Supervised Machine Learning and Natural Language Processing

By Dimitri Kourouniotis

# Developing a Model to Identify Fake Comments

# Exploratory Data Analysis

A sample of 3 million FCC Net Neutrality comments were taken from the 22 million submitted. After importing the data from SQL into a csv file, I cleaned and investigated the data to extract the email, zip code and state details.

My first step was to take a look at the text of the comments and notice the high degree of word pairings.
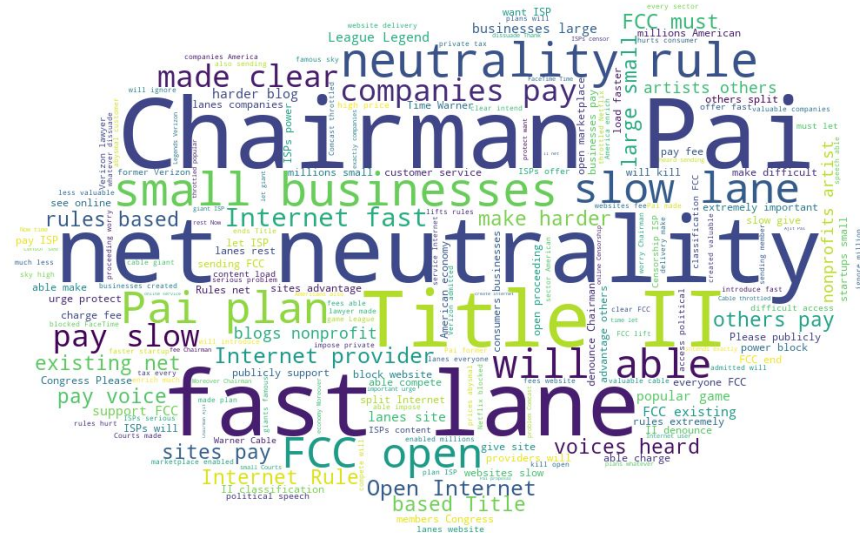
# Emails from Fake Accounts or Domains

Grouping the contact emails by domain we notice that there are over 300,000 'pornhub.com' accounts, as well as over 350,000 from 'hurra.de' domain. I could not find any domain called hurra.de and the pornhub.com accounts had Russian names and fake addresses.



Filings by Email Domain
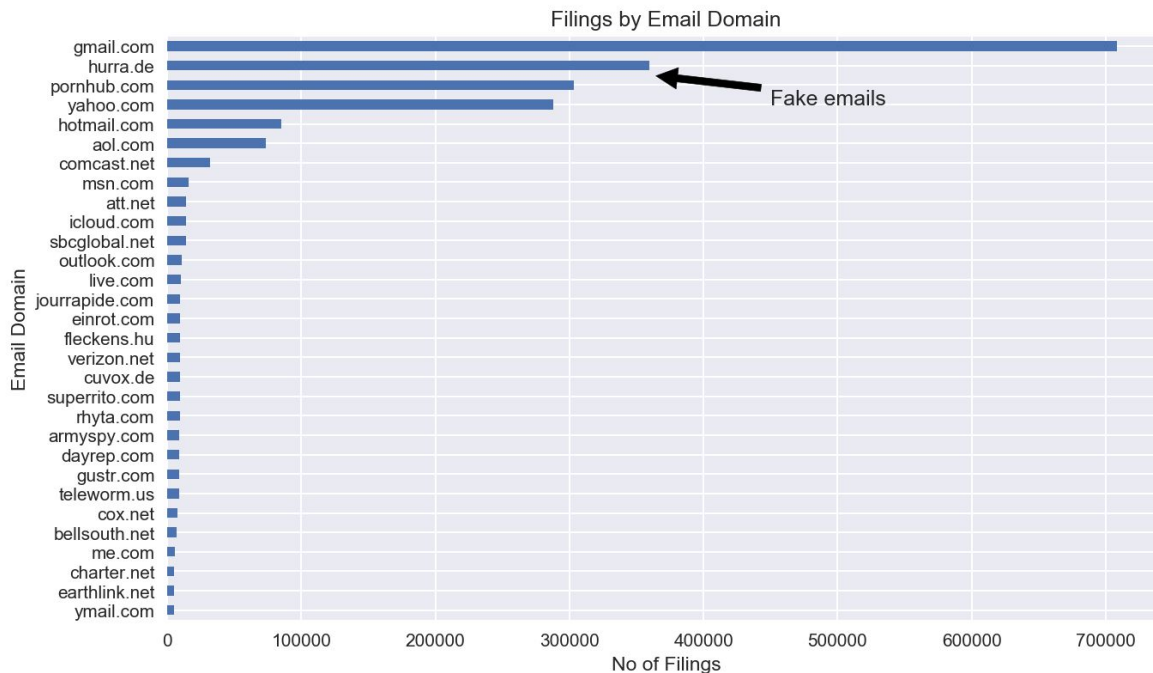
# Submissions per Second

It has already been reported in the press spambots filed hundreds of thousands in the same second. The text of these comments will be a good identifier for Fake or Not.

Here is a diagram of submissions filed in the same second over a 24 hour period starting July 17th 9pm. At one point the peak exceeded 129,000 in the same second (July 17th 8:53:08 pm).



Net Neutrality Comment Submissions
Grouped by Second over 24 hrs

129,719 submissions in the same second
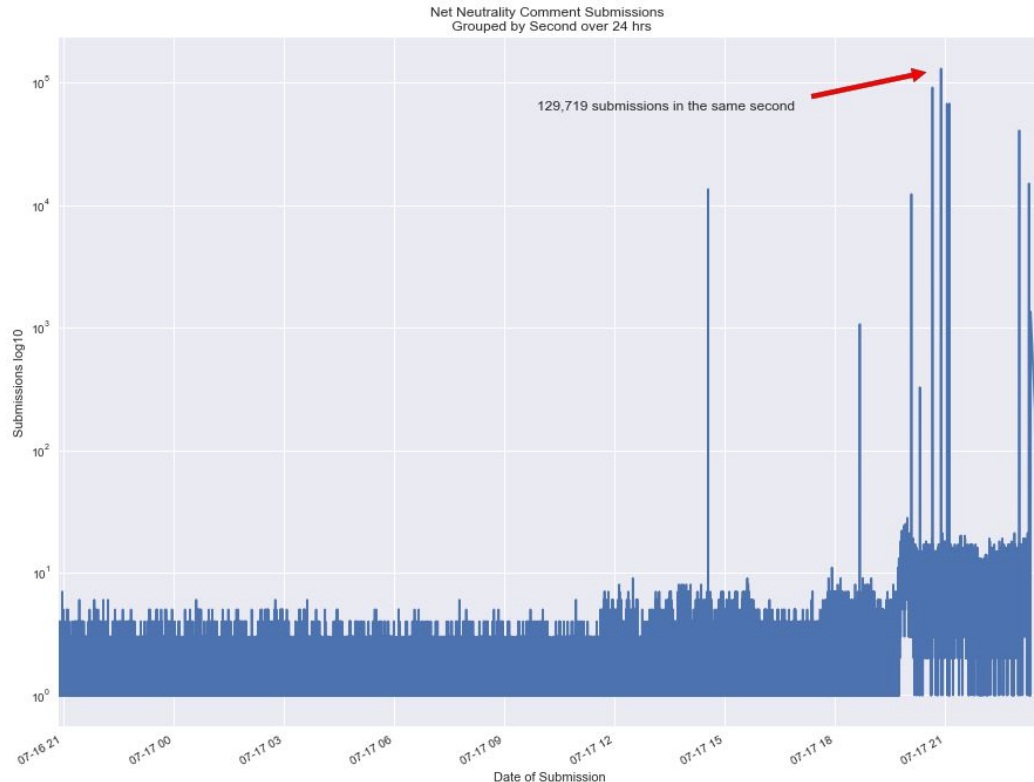
# Variations from the Census

I used census estimates for 2016 population by state to compare them with the filings to see how much variation there was.

The proportion of filings allocated to each state should be similar to the proportion of the US population in that state (and the District of Columbia).

Was this statistically significant?



Number of Filings by State
as they differ from Population Distribution
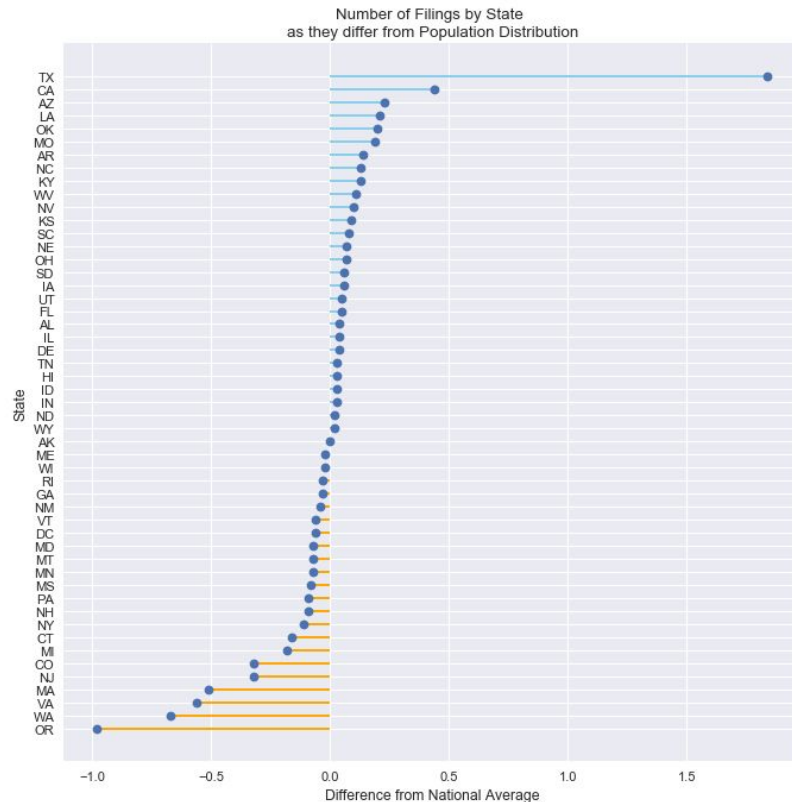
# Statistically Significant from the Population?

Four states have values (Texas, California, Virginia and Oregon) that are well outside the 95% area of the Normal curve as indicated by the horizontal dark red lines either side of the center (2 standard deviations).

From this sample, the Mean is 0, one standard deviation is +/-0.35. Two standard deviations is +/- 0.7.



Distribution of Differences
in FCC Comments filing to State Populations
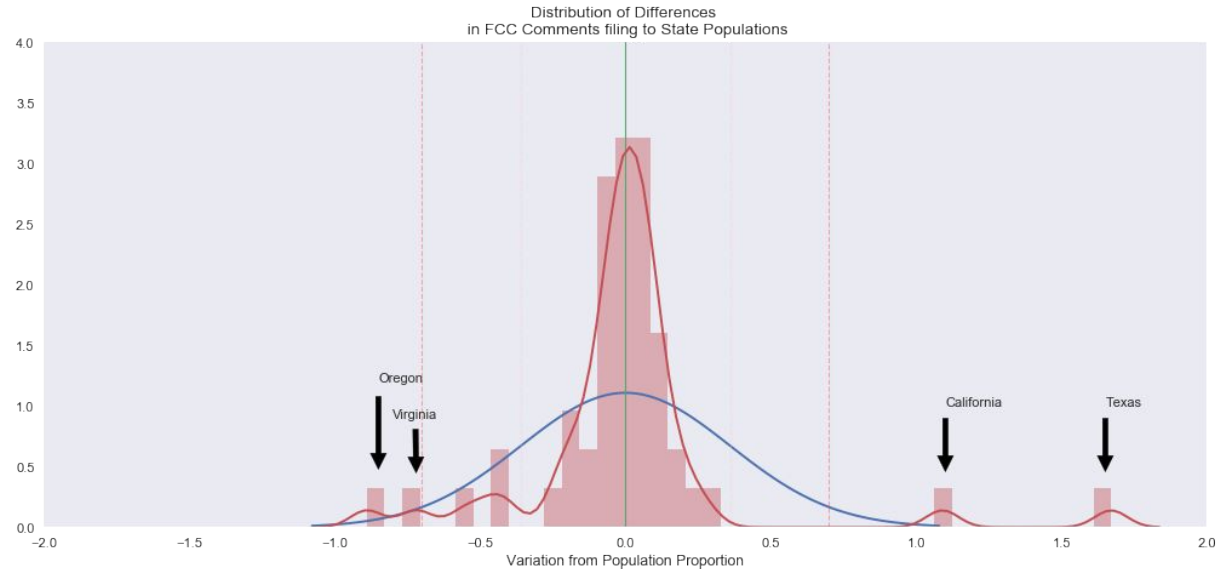
# Where were the Differences?

This diagram illustrates which states, with the more intense hues, had values that were outside the normal distribution.



Net Neutratlity Fillings as Variation of % from State Proportions

% Difference from State
- 0.7-2.0
- 0.36-0.7
- 0.1-0.36
- -0.1-0.1
- -0.36--0.1
- -0.7--0.36
- -2.0--0.7
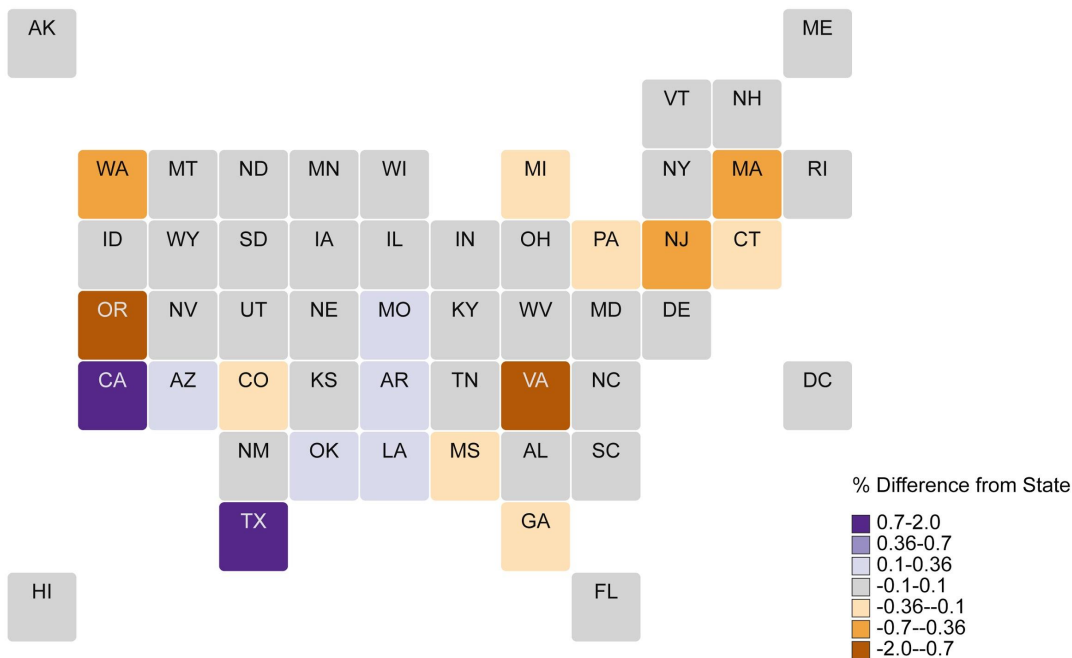
# Sample Text and Feature Labeling

I sampled 500 rows from the dataset labeled them as Fake or Not (1 or 0). Even though the repetition of some of the fake ones is a give away there were other clues, such as missing address details that a local would not make or an email address from pornhub, hurra.de or even an email in all caps with no capitalization in the comments, address etc. Missing value entries in a field rather than blank values, indicating an automated spam bot submitting the form.

| Sample of Fake or Real Text | Fake | Submissions Per Second | Text Length |
|---|---|---|---|
| Net neutrality should NOT be eradicated for th... | 0 | 9 | 205 |
| The unprecedented regulatory power the Obama A... | 1 | 3 | 638 |
| The FCC's Open Internet Rules (net neutrality ... | 1 | 4 | 2311 |
| The FCC's Open Internet Rules (net neutrality ... | 1 | 5 | 2332 |
| The FCC's Open Internet Rules (net neutrality ... | 1 | 129286 | 228 |

# Length of Text as a Feature

Within my sample I also added text comment length and wanted to see if that or submissions per second were also good features. As can be seen on the right, the Fake comments, labeled as '1', have high values for length with the majority of them at 500 and 2300. Whereas the Real comments, labeled as '0', have a text length 500 or less.
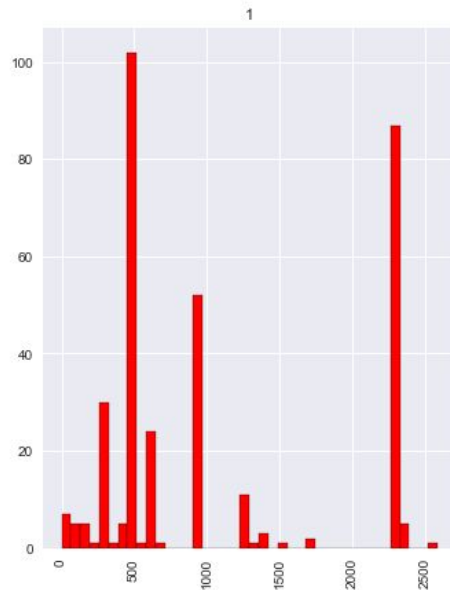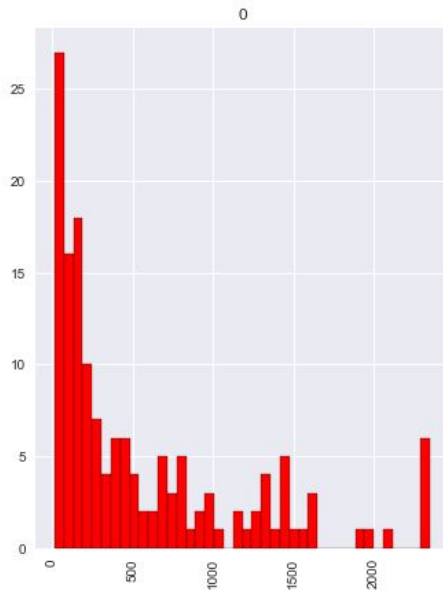
# Classifiers for Supervised Learning

I chose a bag of classifiers to iterate through, with a training size of 60% of the 500 labeled data, and generated the accuracy scores. Initially just on the TFIDF created from the bag of words from the text columns:

SVC
K Nearest Neighbors
Multinomial Naive Bayes
Decision Tree
Logistic Regression
Random Forest
Adaptive Boost
Bagging
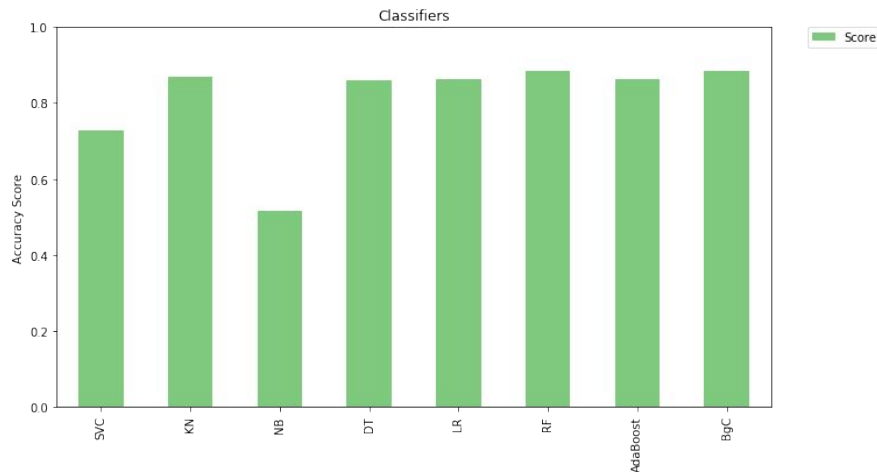


| Classifier | Score |
|------------|----------|
| **SVC** | 0.728643 |
| **KN** | 0.869347 |
| **NB** | 0.517588 |
| **DT** | 0.859296 |
| **LR** | 0.864322 |
| **RF** | 0.884422 |
| **AdaBoost** | 0.864322 |
| **BgC** | 0.884422 |

# Additional Features: Text Length &

Iterating through the TFIDF with the other two features, the Random Forest Classifier gives the best score with the variations and the best result with just the TFIDF and Comment Length.
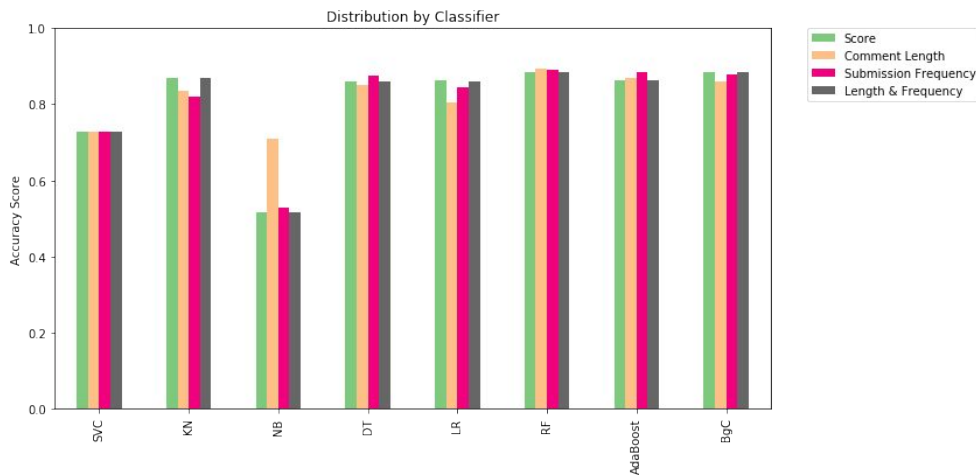


| | Score | Comment Length | Submission Frequency | Length & Frequency |
|---|---|---|---|---|
| **SVC** | 0.728643 | 0.728643 | 0.728643 | 0.728643 |
| **KN** | 0.869347 | 0.834171 | 0.819095 | 0.869347 |
| **NB** | 0.517588 | 0.708543 | 0.527638 | 0.517588 |
| **DT** | 0.859296 | 0.849246 | 0.874372 | 0.859296 |
| **LR** | 0.864322 | 0.804020 | 0.844221 | 0.859296 |
| **RF** | 0.884422 | 0.894472 | 0.889447 | 0.884422 |
| **AdaBoost** | 0.864322 | 0.869347 | 0.884422 | 0.864322 |
| **BgC** | 0.884422 | 0.859296 | 0.879397 | 0.884422 |

# Confusion Matrix from Random Forest

To inspect the results further we can print out the classification result from the confusion matrix:

Classification Report: RFC

|  | precision | recall | f1-score | support |
|---|---|---|---|---|
|  |  |  |  |  |
| 0 (Real) | 0.75 | 0.93 | 0.83 | 54 |
| 1 (Fake) | 0.97 | 0.88 | 0.92 | 145 |
|  |  |  |  |  |
| avg / total | 0.91 | 0.89 | 0.90 | 199 |