Build a Corpus using an API

-Privacy News about US Govt. Agencies

Danish Suri, Krishna Agarwala, Sagar Gupta

Project Phases

- Selecting right API
- Extracting data from API
- Identifying right articles
- Classifying Articles
- Finding accuracy of the data set
- Sentiment Analysis
- Statistically describing data

Why Privacy Corpus for US Government agencies?

- March 2017 had **327** hits for ‘Privacy’ keyword using only the NY Times API.
- Out of these **79** hits were for ‘US’ AND ‘Privacy’ keyword.
- Recently there have been various incidents involving US Government agencies like ‘Snowden revelations’ and ‘Wikileaks snooping’.
- No public database for this which can be used for statistical analysis.
- With the increasing number of privacy incidents involving US Govt. security agencies, there is need for a data set which can help statisticians conduct analysis and research in this topic.

API

New York Times API

The New York Times API allows us to search New York Times articles from Sept. 18, 1851 to today.

We can retrieve headlines, abstracts, lead paragraphs, links to associated multimedia and other article details.

Extracting data from API

Used JavaScript to extract data from API in JSON form.

Code Snippet

```
        function fetchData(page_no)
        {
            var url = "https://api.nytimes.com/svc/search/v2/articlesearch.json";
            var apikey = document.getElementById("apik").value, q=document.getElementById("q").value;
        url += '?' + $.param({
    'api-key': apikey,
    'sort': "newest",
    'page': page_no,
    'q': q
});
$.ajax({
  url: url,
  method: 'GET'
```

Complete Project can be found here: https://github.com/danish20/privacyproject2017

Keywords

That were used to extract data from API

‘CIA AND Privacy’

‘NSA AND Privacy’

‘FBI AND Privacy’

‘DOD AND Privacy’

Total Number of articles for each Keyword

Without Filters

‘Privacy’ - **57865** articles

‘CIA’ - **11619** articles

‘NSA’ - **3221** articles

‘FBI’ - **16926** articles

‘DOD’ - **4890** articles

With Filters:

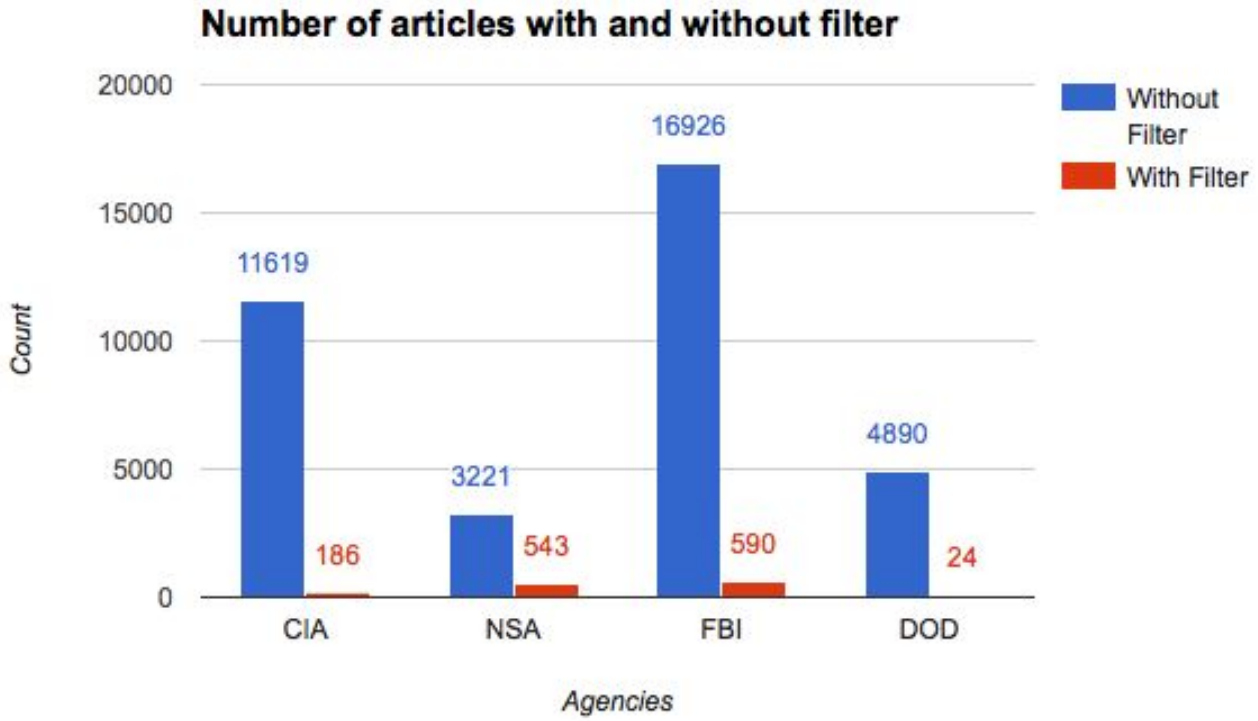
‘CIA AND Privacy’ - **186** articles

‘NSA AND Privacy’ - **543** articles

‘FBI AND Privacy’ - **590** articles

‘DOD AND Privacy’ - **24** articles

Statistical View



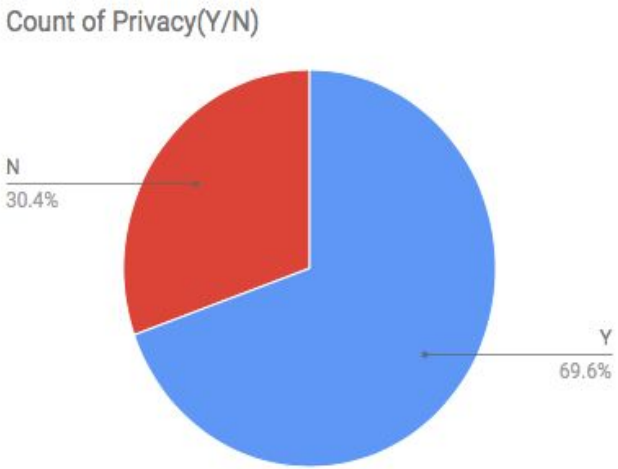
Classifying Articles

After extracting data using API we took a sample of 69 articles for each keyword except for DOD which had only 13 samples and manually labeled our data based on Incident, Date, Privacy related or not, impact and sentiment to have a database to analyse. Following is our data sample:

Tag	Incident	Date	Privacy(Y/N)	Impact	Sentiments	Agency	Keyword
CIA Director Takes Negative Tone on Group Trump Has Praised	CIA-Wikileaks	2017-04-14	Y	World	Negative	CIA	CIA Privacy
CIA Director Calls WikiLeaks 'Hostile Intelligence Service'	CIA-Wikileaks	2017-04-13	Y	World	Negative	CIA	CIA Privacy
Obama Aide Denies Using Intel to Spy on Trump Advisers		2017-04-04	Y	US Government	Negative	CIA+US Govt.	CIA Privacy
Trump Changes Relationship Between White House, Spy Agencies	Trump-Russia	2017-04-01	Y	US Government	Negative	US Govt.	CIA Privacy
Trump's Approach to Intel Agencies Shows Anxiety, Distrust	Trump-Russia	2017-03-31	Y	US Government	Negative	US Govt.	CIA Privacy
Apple: Software Flaws in Latest WikiLeaks Docs Are All Fixed	CIA-Wikileaks	2017-03-24	Y	World	Positive	Apple+CIA	CIA Privacy
Trump's Russian Imbroglio Prompts Republican Rethink on Surveilla	Trump-Russia	2017-03-20	Y	World	Neutral	US + Russia Gov	CIA Privacy
Yahoo Breach Indictments May Shed Light on Other Hacks	Yahoo Breach	2017-03-16	Y	World	Positive	CIA+Yahoo	CIA Privacy
Russian Agents, Hackers Charged in Massive Yahoo Breach	Yahoo Breach	2017-03-15	Y	World	Negative	Russia+CIA+Yah	CIA Privacy

Statistically Describing Findings

CIA



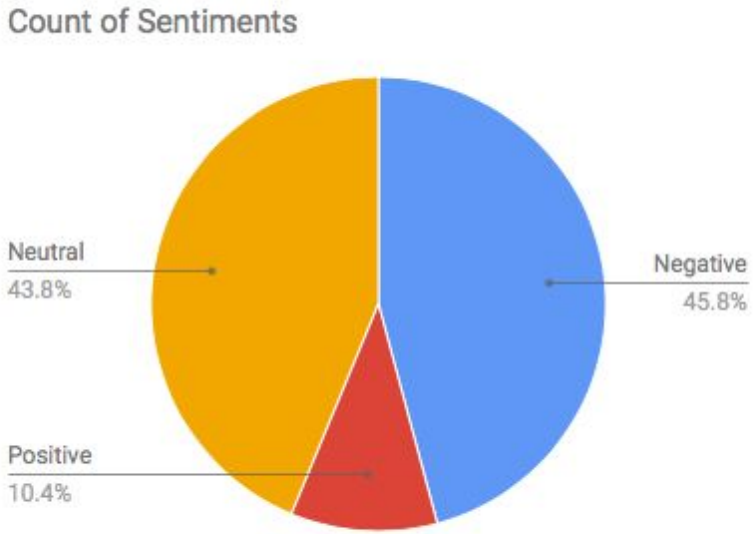
Sample article = 69

True Positive = 48

False Positives = 21

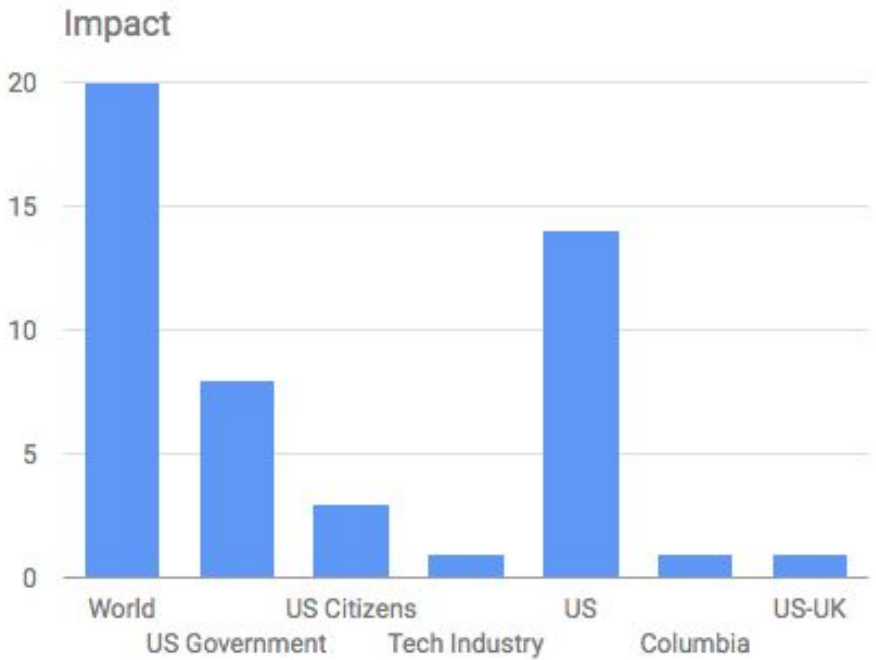
Precision = 69.6%

Sentiment Analysis

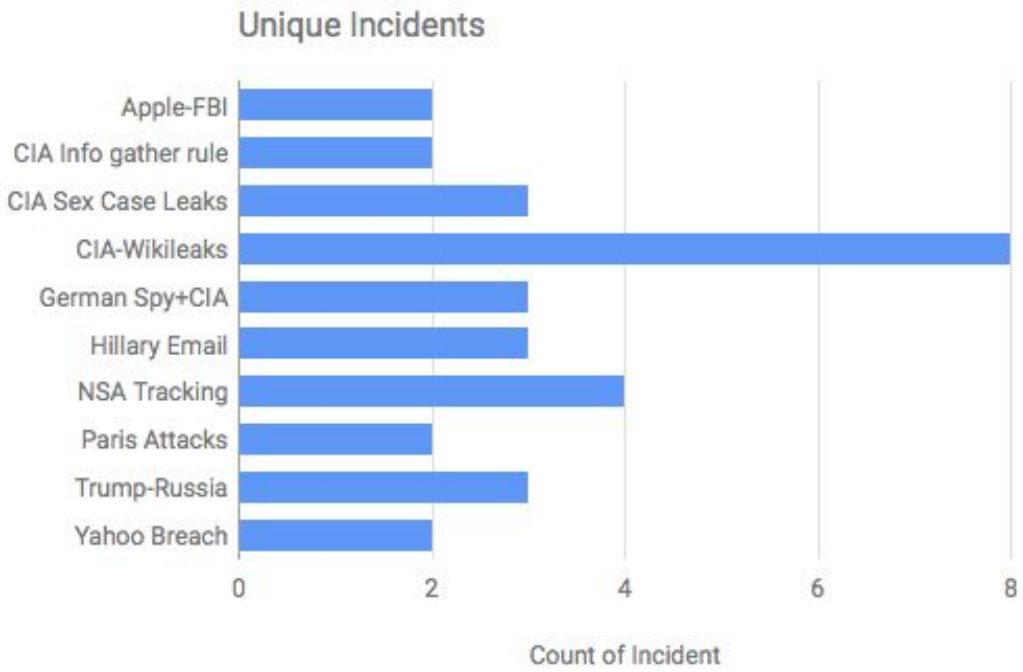


Sentiment	Count
Positive	5
Negative	22
Neutral	21

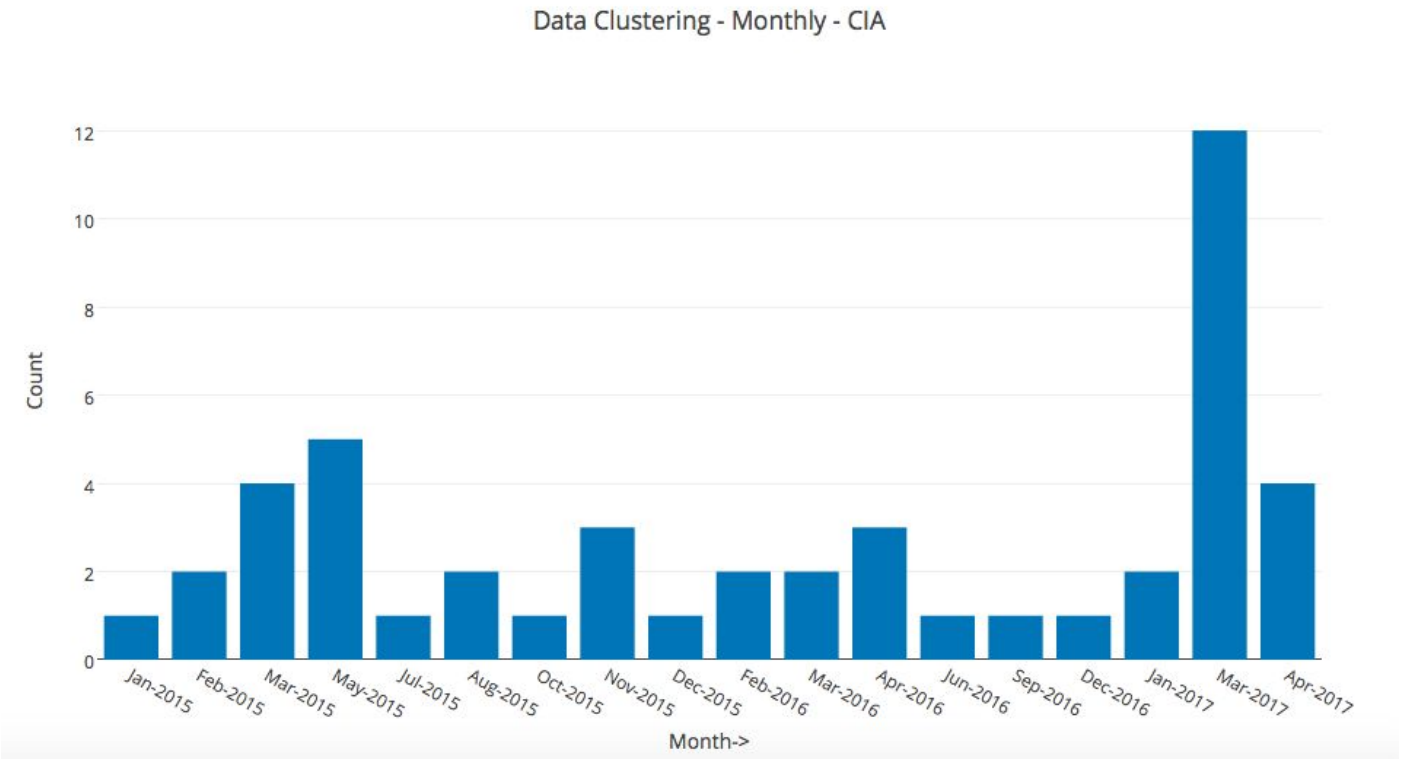
Impact


Impact
20
15
10
5
0
World
US Government
US Citizens
Tech Industry
US
Columbia
US-UK

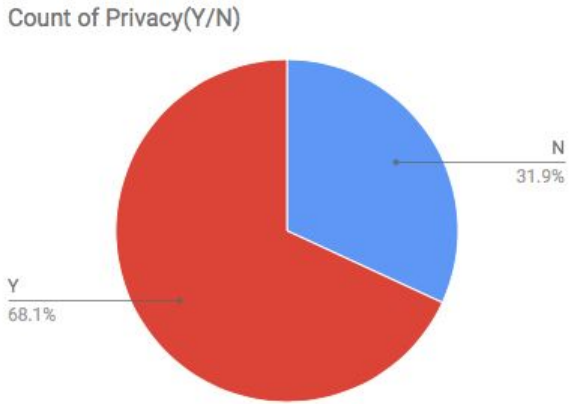
Unique Events



Data Clustering - Monthly



NSA



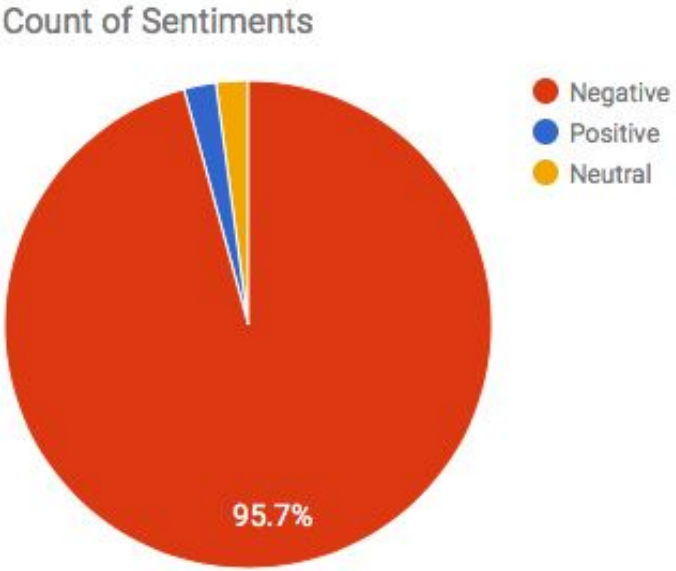
Sample article = 69

True Positive = 47

False Positives = 22

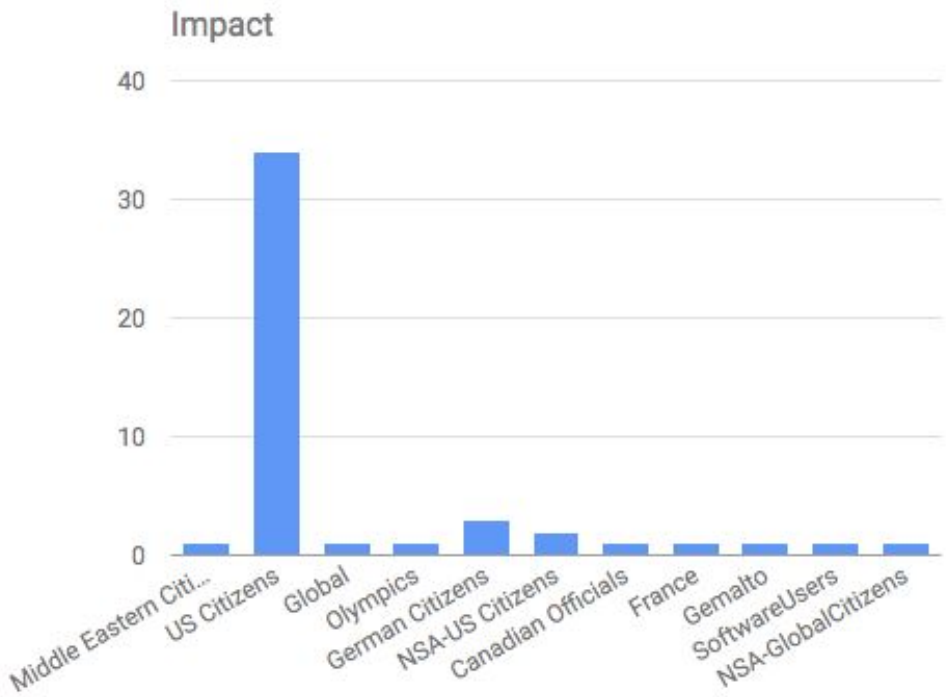
Precision = 68.1%

Sentiment Analysis

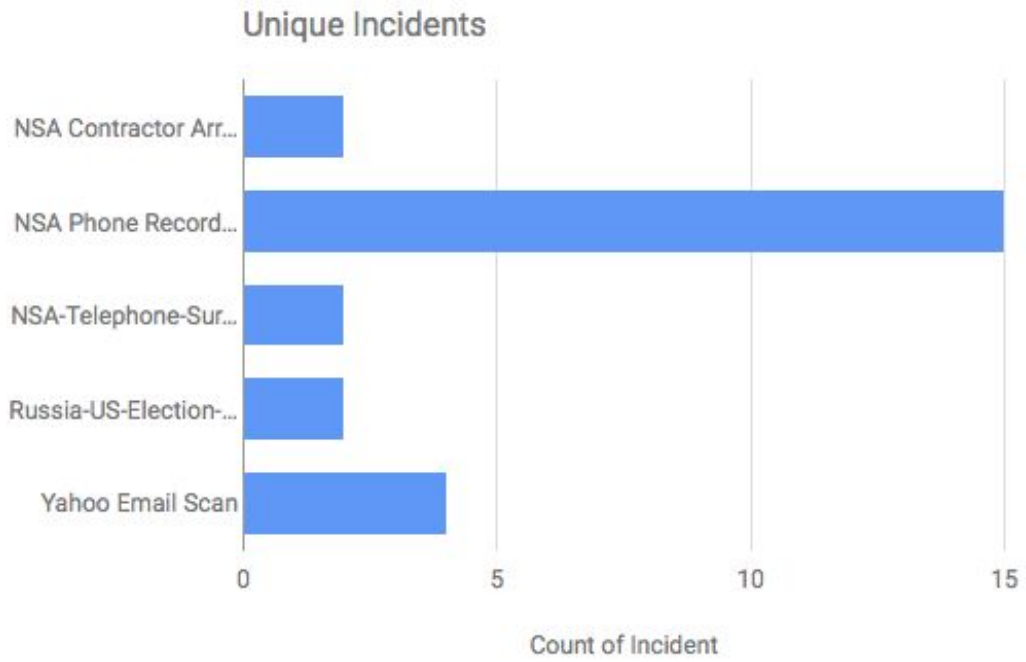


Sentiment	Count
Positive	1
Negative	45
Neutral	1

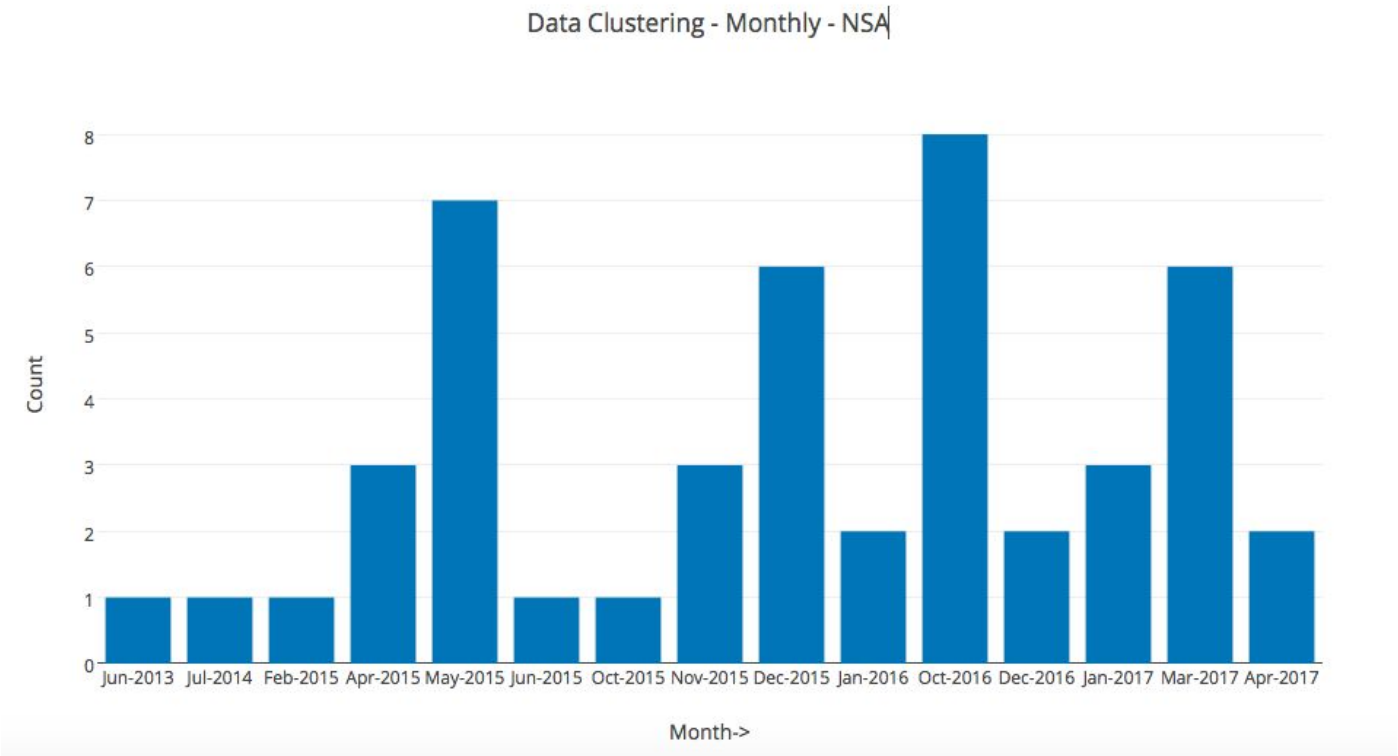
Impact


Impact
40
30
20
10
0
Middle Eastern Citi...
US Citizens
Global
Olympics
German Citizens
NSA-US Citizens
Canadian Officials
France
Gemalto
SoftwareUsers
NSA-GlobalCitizens

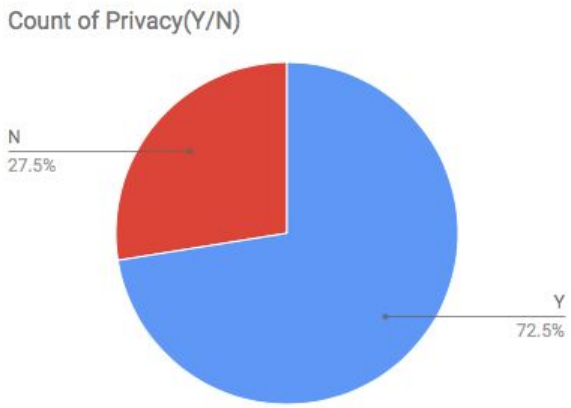
Unique Events



Data Clustering - Monthly



FBI



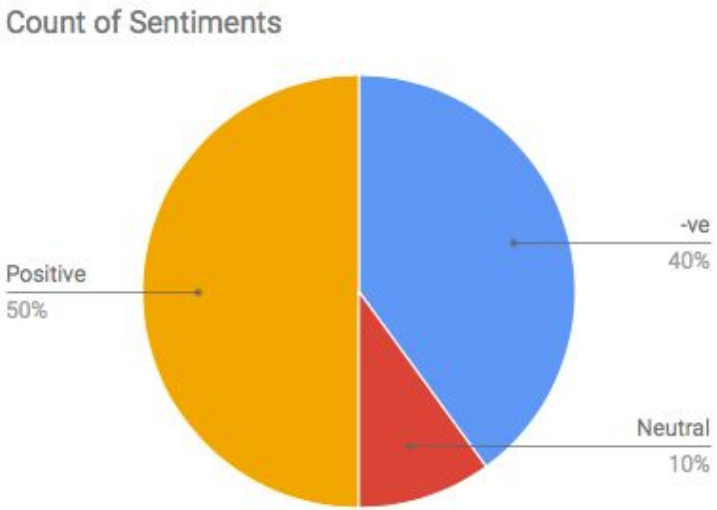
Sample article = 69

True Positive = 50

False Positives = 19

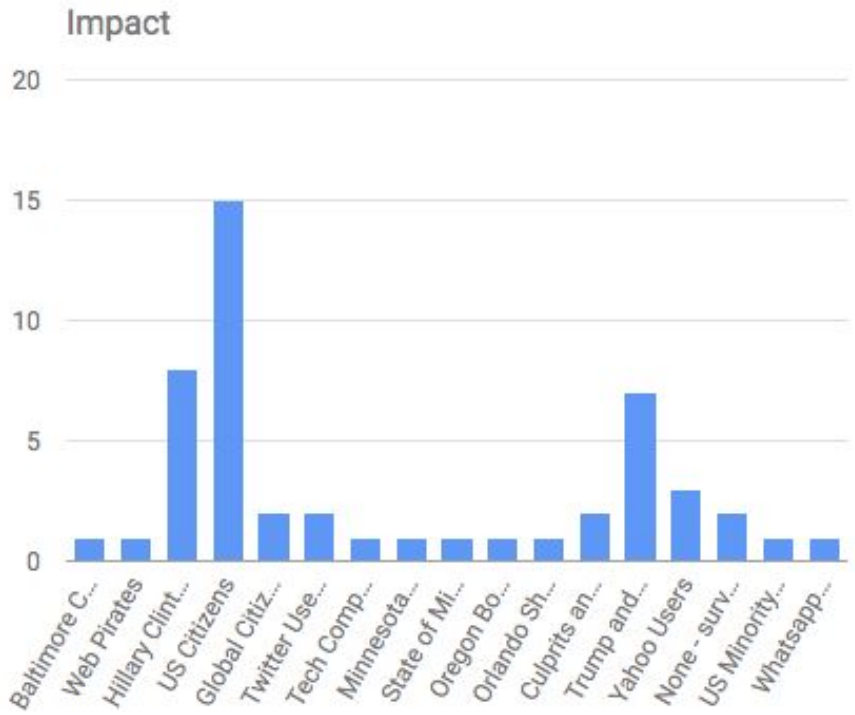
Precision = 72.4%

Sentiment Analysis

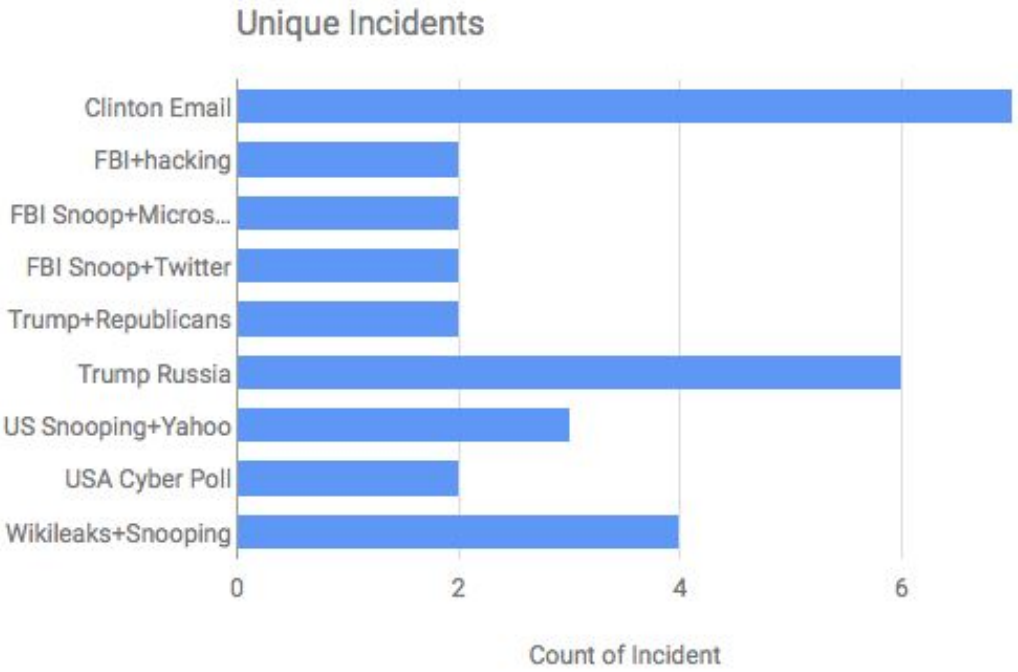


Sentiment	Count
Positive	25
Negative	20
Neutral	5

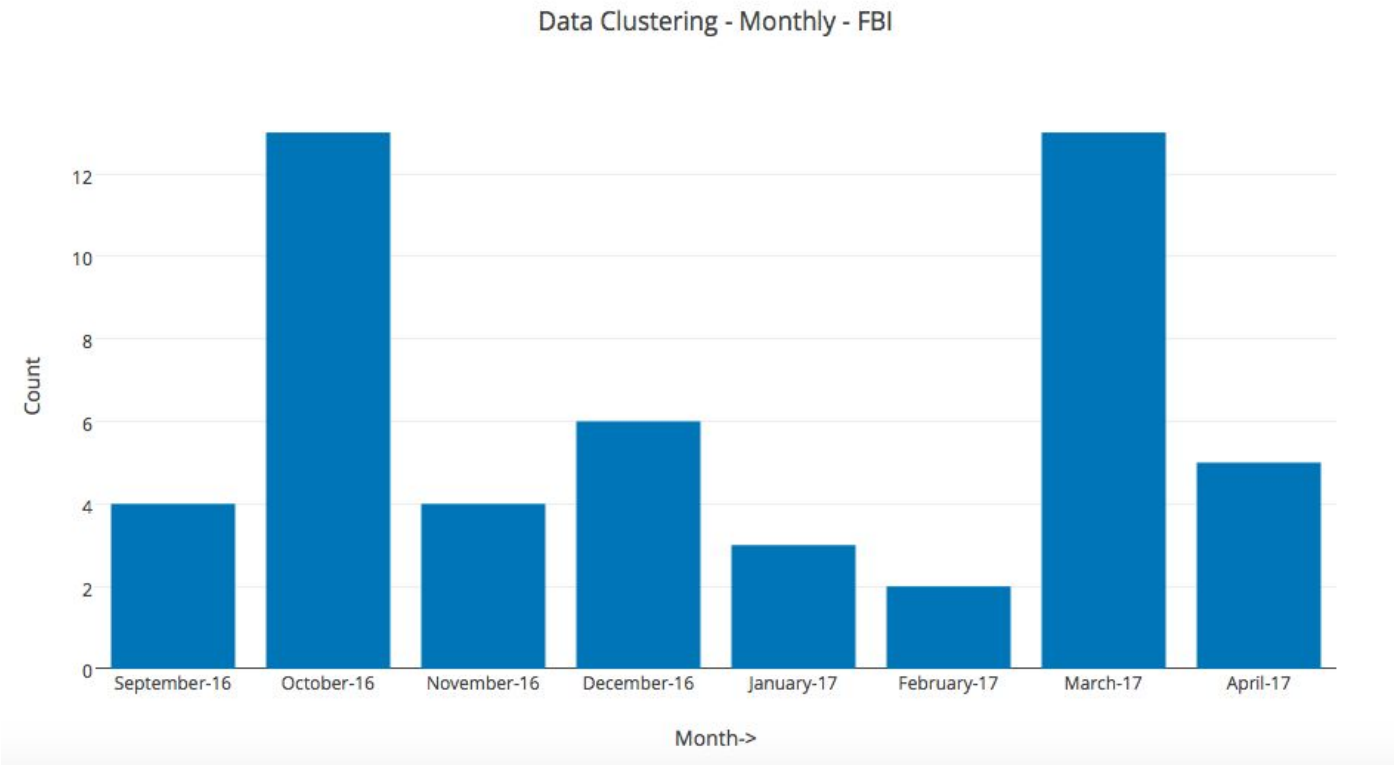
Impact


Impact
20
15
10
5
0
Baltimore C...
Web Pirates
Hillary Clint...
US Citizens
Global Citiz...
Twitter Use...
Tech Comp...
Minnesota...
State of Mi...
Oregon Bo...
Orlando Sh...
Culprits an...
Trump and...
Yahoo Users
None - surv...
US Minority...
Whatsapp...

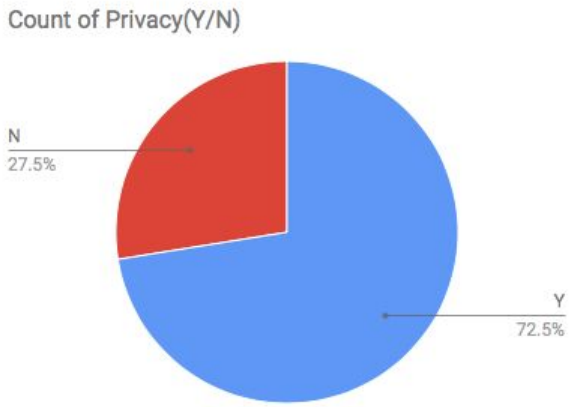
Unique Events



Data Clustering - Monthly



DOD



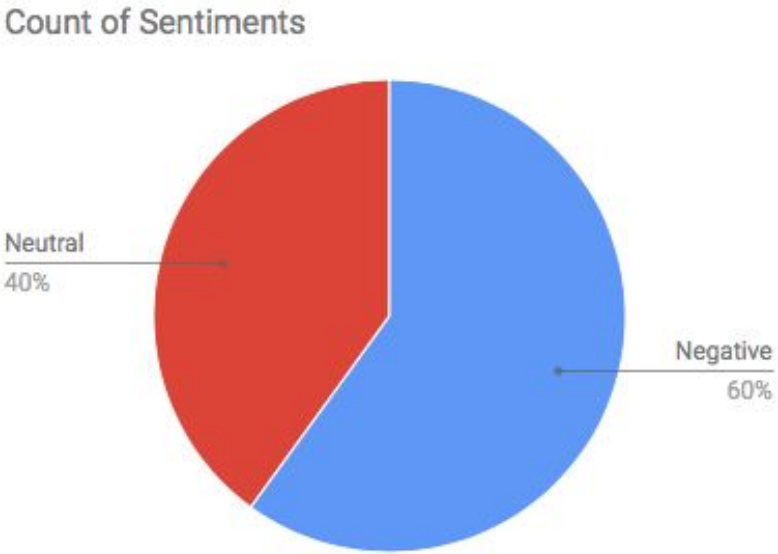
Sample article = 13

True Positive = 10

False Positives = 3

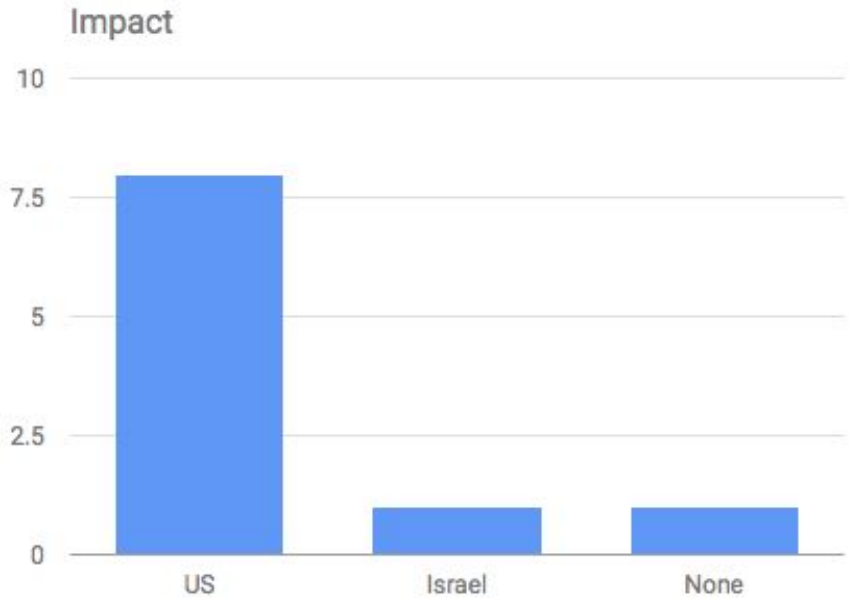
Precision = 76.9%

Sentiment Analysis



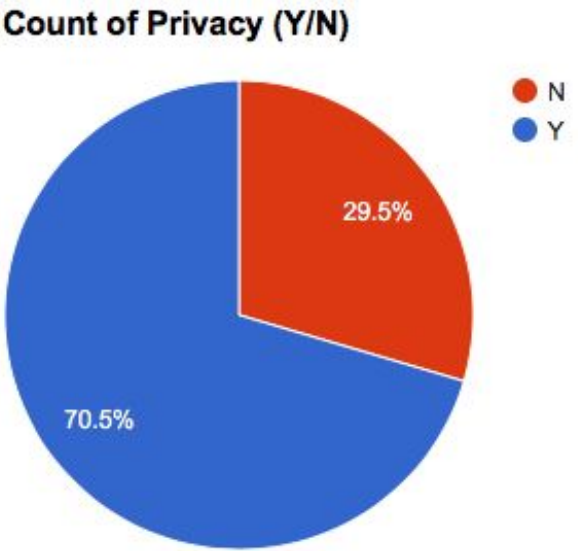
Sentiment	Count
Positive	0
Negative	6
Neutral	4

Impact



Overall Data Analysis

Precision



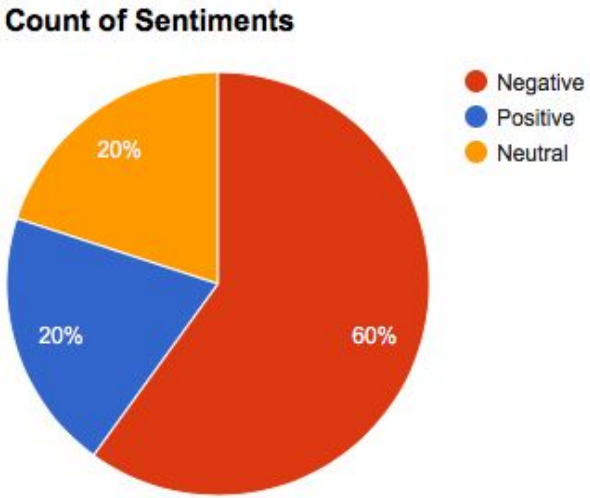
Total Sample= 220

True Positive = 155

False Positives = 65

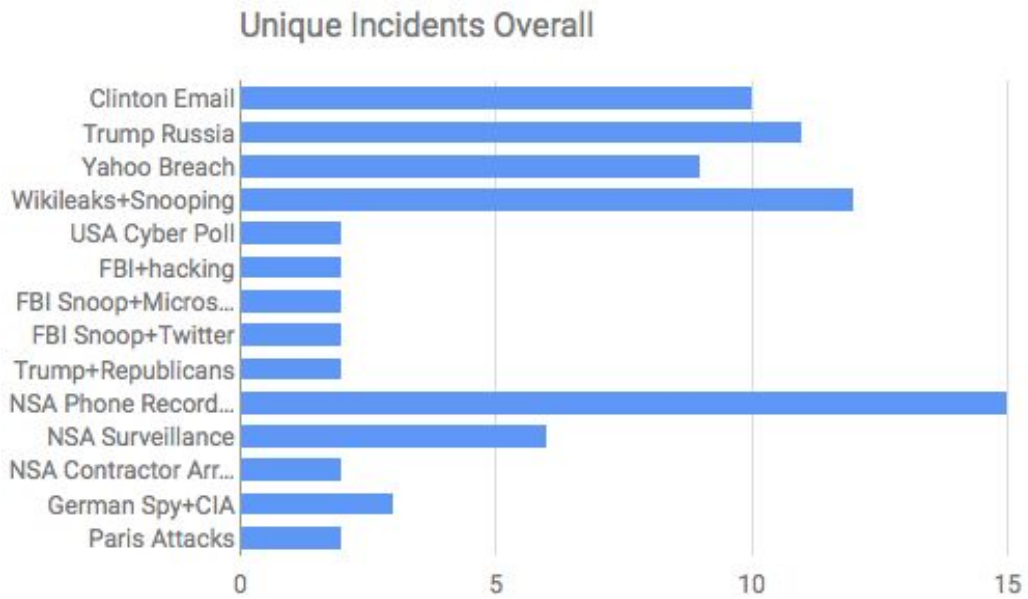
Precision = 70.5%

Sentiment Analysis

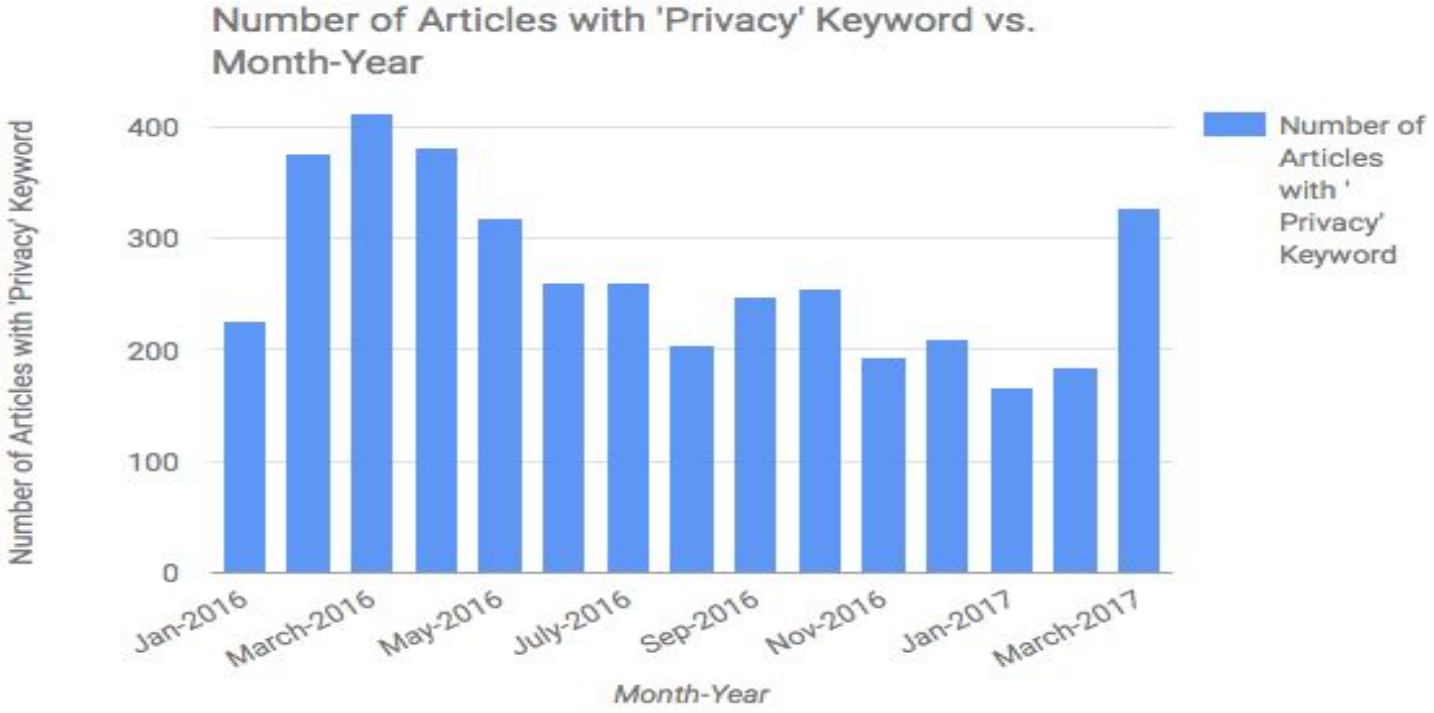


Sentiment	Count
Positive	44
Negative	**132**
Neutral	44

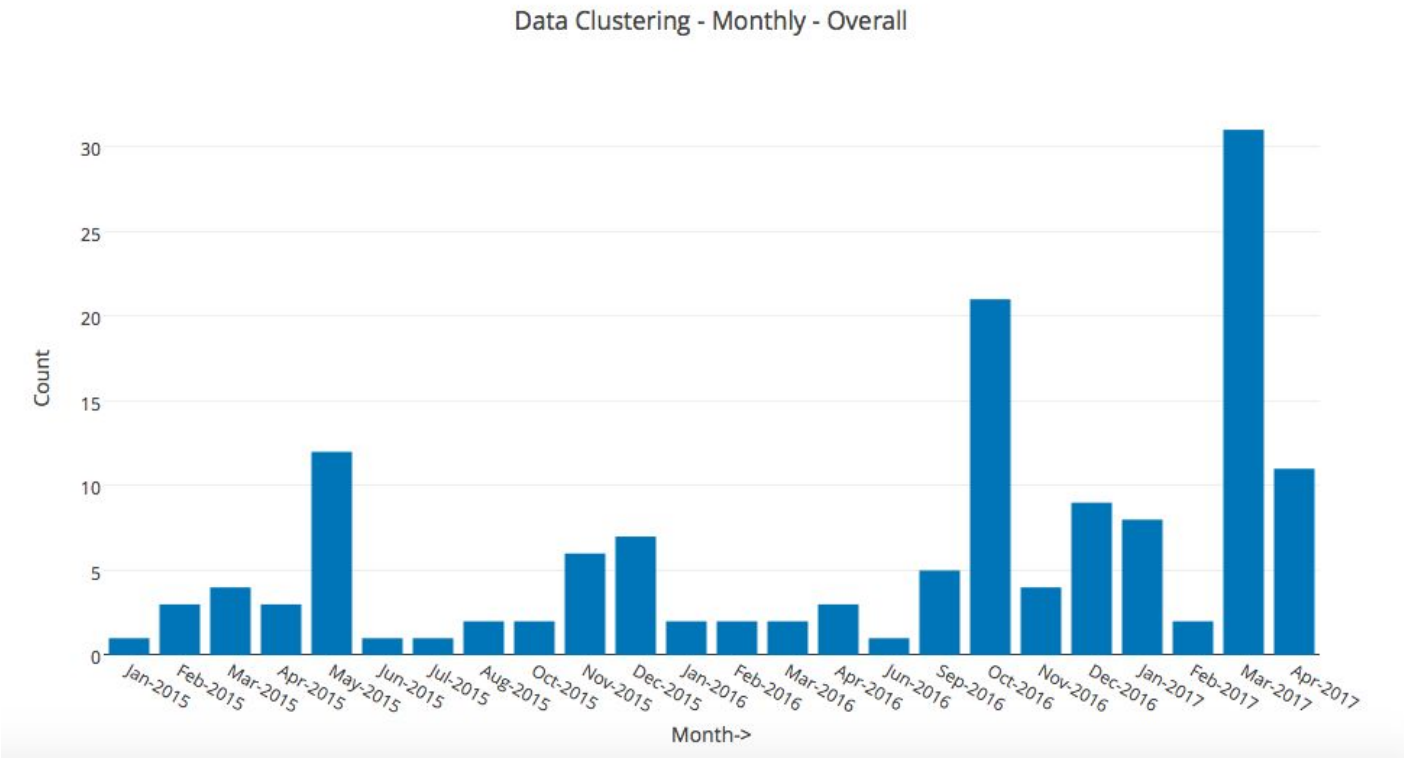
Unique Incidents Clustering



Data Clustering - Monthly for ‘Privacy’ Keyword



Data Clustering - Monthly



Strengths And Weaknesses

- Strengths
  - There is a plenty of data and news available related to US Govt. agencies.
  - Simple keyword searches like ‘FBI+Privacy’ can result in number of articles.
  - NYT API was fairly simple and provided data in JSON format which was easy to analyze and represent.

- Weaknesses
  - Only NY Times provide a well written API for US news.
  - **Missing Data:** There could be a plenty of news articles which were not covered by NYT but other news papers related to privacy news involving US Govt. agencies.
  - **Noise**: Analysing data based on keywords is very difficult as articles with keywords related to government agencies and ‘privacy’ may represent some positive news.

Work Distribution

- Danish
  - Analysing various news agencies' API to fetch data and choosing one for the project.
  - Evaluating the chosen API and reading its documentation.
  - Summarizing the final data and performing analysis.
- Sagar
  - Use the chosen API to gather a well represented set of data which will be used as 'Training Data'.
  - Determining various tags to query the API.
  - Summarizing the final data and performing analysis.
- Krishna
  - Manually analyzing the training data to determining accuracy.
  - Calculating precision of the final data obtained.
  - Summarizing the final data and performing analysis.


Thank You