SENTIMENT ANALYSIS ON TWEETS OF AADHAAR

Introduction

Twitter sentiment analysis

Eventhough one lives in the fear of machines attaining the capability of emotions, there is always a curiosity of what it can accomplish. Sentiment analysis is one such tool that was created to tackle the eternal problem of emotional understanding for machines. It looks at the opinion or feeling of a certain text.

Sentiment analysis is the process of determining whether a piece of writing is positive, negative or neutral. All words and phrases that imply positive or negative sentiment are taken and rules are applied that consider how context might affect the tone of the content. These carefully crafted rules then help discovering the polarity of the data. It is extremely useful in monitoring social media as it allows one to gain and overview of the wider public opinion behind certain topics.

Sentiment analysis has its limitations as it cannot be used as 100% accurate in any given scenarios. But this can be overcomed when it's agreed upon that all human expressions cannot fit into three categories. Also, the insights that can be gained from a large dataset will overshadow these concerns of reliability at a granular level.

As, human preferences are practically unpredictable but with data being freely available, data scientists can test hypothesis using the ultimate psychological tool - twitter. Twitter is a treasure trove of sentiment. People around the world, output thousands of reactions and opinions on every topic under the sun every second of every day. It’s like one big psychological database that's constantly being updated and we can use it to analyse millions of technological snippets in seconds with the power of machine learning.

Through this project I’m trying to process a python script that uses twitter tweets to understand how people are feeling about a certain topic.

Topic - Aadhaar

Aadhaar is a 12-digit unique identification number mandated to all residents of India since 2016 by the statutory authority established by the government of India called unique identification authority of india (UIDAI). The number is linked to resident’s basic demographic and biometric information which are then stored in the centralised database.

Being the largest biometric ID in the world, there have been growing concerns regarding it. The main concerns being regarding “privacy, potential for surveillance and exclusion of eligible beneficiaries from welfare schemes from the leveraging of Aadhaar based systems. Due to which, the Aadhar project validity is being challenged in the supreme court of India, till date.

It has been a hot topic of discussion for several years now and with growing experiences and information into the public there has been much revolt against it. Most often used platform to express one's opinion is the social media and twitter is living through these varied discussions.

Research question

Through this project I would like to understand the mass opinion about Aadhaar using twitter tweets and executing a sentiment analysis on the corpus. This will check the polarity of the topic and visualize the same.

Database

The dataset was created using the design prototype of documenting the now (docsnow). It consists of 10,000 tweets with the words aadhaar. The data is open source and easily available to download in the csv format.

Link: http://app.docnow.io/summary/201712291119-bd4448/

Processing the data

Cleaning the data - using open refine and other

1) Imported the data onto Open Refine and deleted all the unwanted columns and kept only the tweet text and the language column.
2) Using the text facet, kept only the english language tweets as the code reads only the particular language.
3) Used the common transformations such as trimming leading and trailing whitespaces, removing blank cells etc.
4) Removed symbols such as RT (retweet) from all the tweet data as it was unnecessary for the analysis.
5) Avoided clustering and removing duplicate cells as it would hinder the mass opinion of the people. Therefore same tweets exist but by different people hence counted in the large chunk.
6) Removed all the non-ascii characters from the data using Diacritics remover (http://utils.paranoiaworks.org/diacriticsremover/) and converted them into white spaces.
7) Exported the data into csv format before I could import it into the python code.

By end of the cleaning, the tweets had reduced to 3838 before I could process it.

Code

The python code used was the VADER (Valence Aware Dictionary and sEntiment Reasoner) which is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media, and works well on texts from other domains too. The VaderSentiment is fully open sourced under the MIT license. (https://github.com/cjhutto/vaderSentiment)

Steps involved:

1) Installed python 3.6.4, as the VaderSentiment code works only in it.
2) Installed pip
3) Installed VaderSentiment and its dependencies.
4) Used the code and manually entered the corpus into the code (as eventhough I managed to import data using a csv file but failed to make it read line by line as the data cleaning wasn't completely up to the mark).
5) Processed the code and got a fullout text with the positive, negative, neutral and compound score for each tweet.
6) Used the output in sublime text and added commas between the values and saved it as a csv file.
7) After separating into columns generated visualisations for analysis.

Code Output

The output for the dataset was scores of positive, neutral, negative and compound.

- The compound score is calculated by summing the valence scores of each word in the lexicon, adjusted according to the rules and normalised to be between -1 and +1. This is a unidimensional measure of sentiment for a given data.
- It has typical threshold values (positive sentiment: compound score >= 0.5| neutral sentiment: (compound score > -0.5) and (compound score < 0.5) | negative sentiment: compound score <= -0.5)
- The positive, neutral and negative scores are ratios for the proportions of data in each tweet.

Visualisation and Analysis

For the purpose of visualisation, I chose to do a uni-dimensional analysis of compounded score. The polarity of the tweets can be analysed through this score. Following are two representations of the compounded scores.

1) **The distribution curve of the compounded score.**

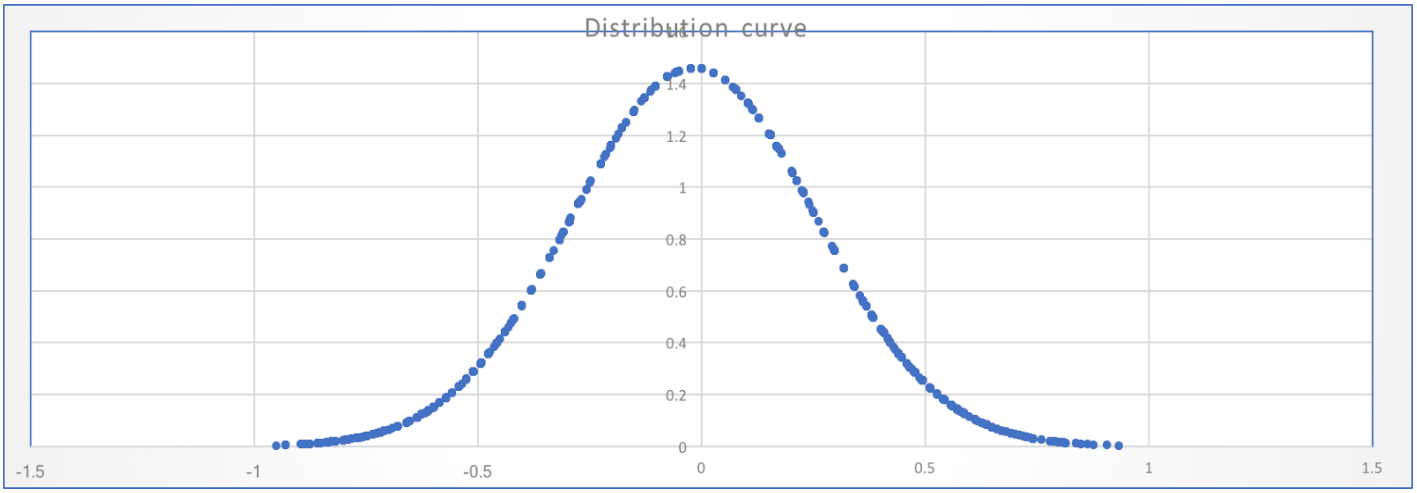

Figure 1

The output data contained a compounded score for each tweet. The mean and standard deviation (SD) of these scores was evaluated. The mean score was -0.01770206 and the SD was 0.273338822. Then the distribution was calculated using Microsoft excel sheet. The distribution curve was then plotted. The same is shown in Figure 1 above.

From the curve we infer that the compounded scores follow a normal distribution pattern. This shows that there are no dramatic variations in data. However, the mean is slightly less than zero. The spread of the curve shows that more than 1SD lies between -0.5 and 0.5, which means that majority of the tweets (more than 68%) are neutral in nature.

2) **Doughnut chart of the compounded score**

The compounded scores were divided into three categories as under:

Category	Polarity	Score
1	Negative	< -0.5
2	Neutral	-0.5 - 0.5
3	Positive	> 0.5

The count of scores in each category was found. Since all scores lied in one of the three categories and none lied in more than one category, a percentage for each category could be found easily. This percentage directly corresponds with the polarity of the tweets. The mapping is as given under.

Next, a doughnut chart was plotted to better visualise the polarity of the tweets. As seen from Figure 2, 89% of tweets were neutral in nature. About 7% of the tweets were negative and 4% were positive. This indicates that overall sentiment about Aadhar is of the neutral kind. However, the percentage of people who had tweeted negatively is greater than the people who tweet positively.

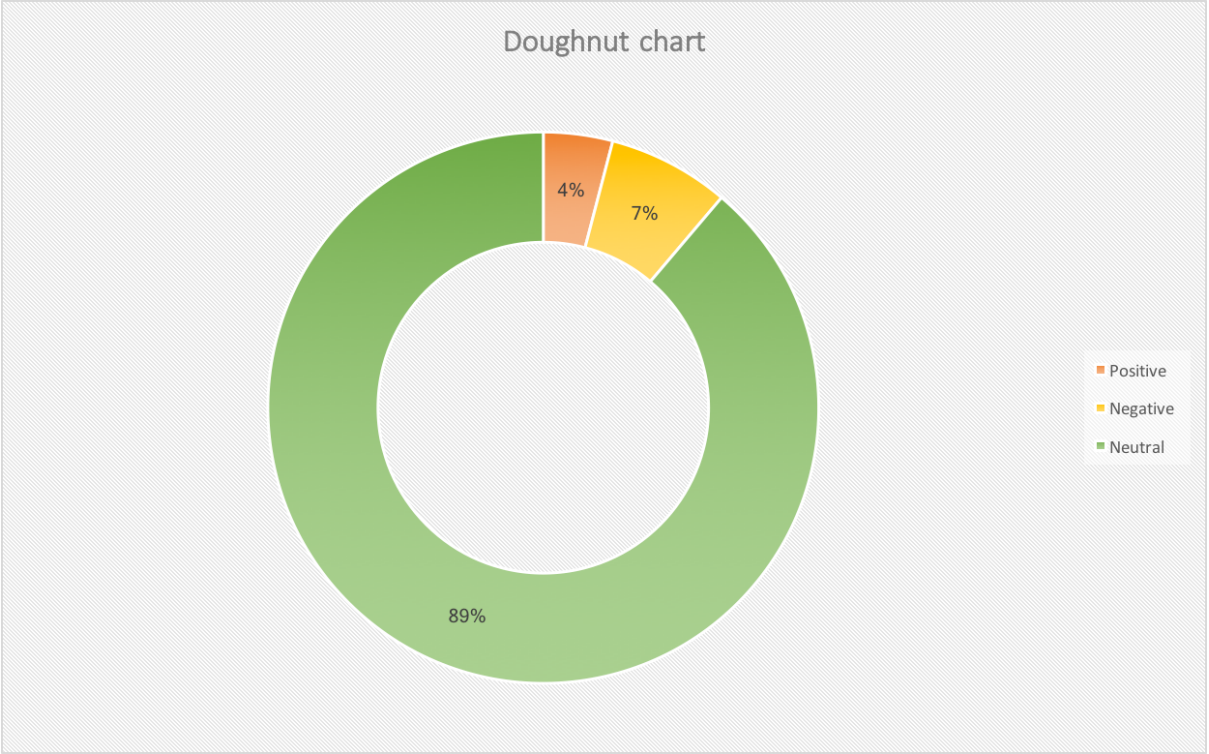

Figure 2

Conclusion

Aadhaar has been a growing concern for the people of India for many years now and it continues to be until the satisfaction criteria are reached. Having the social media platform to voice out one's opinion has allowed the 'makers' or even the government to understand the needs of such a huge population. Tools such as sentiment analysis help visualise this mass of live data and understand the concerns regarding mandated rules such as Aadhaar. In this case, the result shows a huge neutral population and not positive which itself speaks much of people's opinion on this mandated rule. The slight deviation towards negative would probably change with increasing the amount of database. Furthering the analysis by applying multi dimensional analysis would surely pinpoint the cause for such high neutrality. Therefore, the scope for twitter sentiment analysis is very promising.