Understanding People Opinion about Aadhaar Using Sentiment Analysis on Twitter

Mahesh G Huddar[1*], Pradeep J Gorawade[2], Ravisagar S Joshi[3], Hiragond Nikhilkumar M [4], Manjunath Poleshi C[5]

[1]Assistant Professor*
Dept. of Computer Science and Engineering
Hirasugar Institute of Technology, Nidasoshi, Belagavi
mailtomgh1@gmail.com

[2,3,4,5]Graduate Students
Dept. of Computer Science and Engineering
Hirasugar Institute of Technology, Nidasoshi, Belagavi

Abstract: 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 centralized database. Being the largest biometric ID in the world, there have been growing concerns regarding it. The main concerns areprivacy, potential for surveillance and exclusion of eligible beneficiaries from welfare schemes from the leveraging of Aadhaar based systems. Due to which, the Aadhaar 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. The main objective of this work is to show how sentiment analysis can help in understanding people opinion about Aadhaar.

Keywords: Aadhaar Sentiment Analysis, Machine Learning, Bag-of-Words, Bayesian Networks

1. Introduction

Data mining is a process of mined valuable data from a large set of data. Several analysis tools of data mining (like, clustering, classification, regression etc,) can be used for sentiment analysis task [1][2]. Even though 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 used to post 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.

2. Related Work

Sentiment analysis is addressed using many approaches. Some of the work is discussed in this section.

a. Mori Rimon[3]classified sentiment using the keyword based approach. They worked on identifying keywords basically adjectives which indicate the sentiment. Such indicators can be prepared manually or derived from Word net.

b. Alec co[4] used different machine learning algorithms such as Naïve Bayes’, Support vector machine and maximum entropy.

c. Janice M. Weibe[5]performed document and sentence level classification. He fetched review data from different product destinations such as automobiles, banks, movies and travel. He classified the words into positive and negative categories. He then calculated the overall positive or negative score for the text. If the number of positive words is more than negative then the document is considered positive otherwise negative.

d. Jalaj S. Modha ,Gayatri S. Pandi and Sandip J. Modha[6] worked on techniques of handling both subjective as well as objective unstructured data.

e. Theresa Wilson, JanyceWiebe and Paul Hoffman[7] worked on a new approach on sentiment analysis by first determining whether an expression is neutral or - Tracking users and non-users opinions and ratings on products and services. - Monitoring issues confronting the company so as polar and then disambiguates the polarity of the polar expression. With this approach the system is able to automatically identify the contextual polarity for a large subset of sentiment expressions, hence achieving results which are better than baseline.

3. Methodology

Twitter is a micro blogging service usually used as an instant communication platform. The capacity to provide information in real time stimulated many companies to use this service to understand their consumers. 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 on that content. These carefully crafted rules can help in discovering the polarity of the data. We are developing a system which can determine the opinions about Aadhaar of peoples and determining whether it is positive or negative. We are extracting the data from Twitter social media, after extracting the data we are Pre-processing the data in which related and unrelated tweets are found out. All unrelated tweets are considered to be outliers. We are going to propose system, which has various important parts as Data Extraction, pre-processing of extracted data and classification.

- **Data Extraction:** There are various tools are available for extracting the data from heterogeneous sources. We used twitter API to extract aadhaar related tweets form twitter.
- **Data Pre-Processing:**
  - Imported the data onto Open Refine and deleted all the unwanted columns and kept only the tweet text and the language column.
  - Using the text facet, kept only the English language tweets as the code reads only the particular language.
  - Used the common transformations such as trimming leading and trailing whitespaces, removing blank cells etc.
  - Removed symbols such as RT (retweet) from all the tweet data as it was unnecessary for the analysis.
  - 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.
  - Removed all the non-ascii characters from the data using Diacritics remover[8] and converted them into white spaces.
  - Exported the data into csv format that is input to the classification step.
- **Classification:** After Pre-processing the data, the collected text is classified using Bayesian classifier.Based on this analysis, we predict the given tweet is positive, negative or neutral.

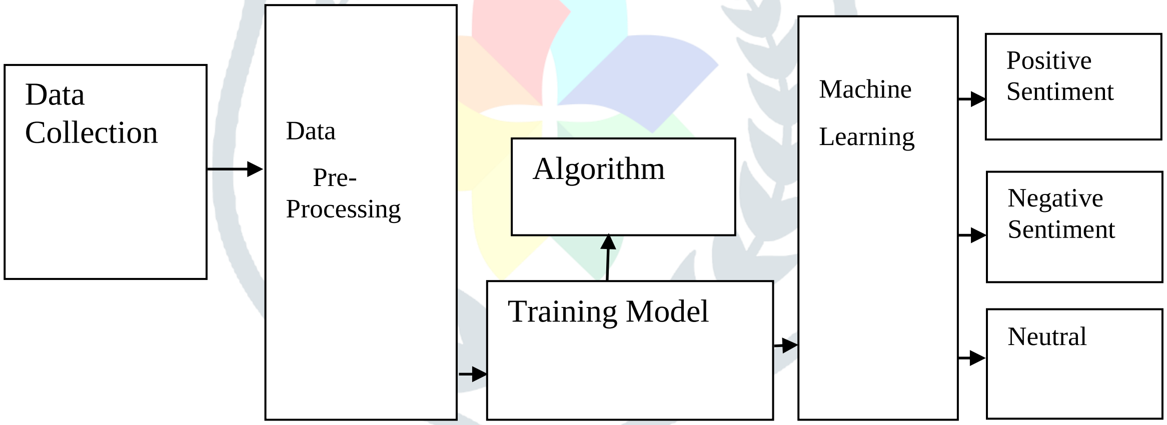

Figure 1. Data Flow Diagram

Table 1. Sample tweets extracted from Twitter

SAMPLE TWEETS
Edward Snowden Verified account @Snowden Jan 21 MoreRarely do former intel chiefs and I agree, but the head of India's RAW writes **#Aadhaar** is being abused by banks, telcos, and transport not to police entitlements, but as a proxy for identity–an improper gate to service. Such demands must be criminalized
#Aadhaar is an identifier, not a profiling tool. Aadhaar database does not keep any information about bank accounts, shares, mutual funds, property details, health records, family details, religion, caste, education etc. #AadhaarMythBuster

4. Implementation

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

Bayesian network classifiers are a popular supervised classification paradigm. A well-known Bayesian network classifier is the Naïve Bayes' classifier is a probabilistic classifier based on the Bayes' theorem, considering Naïve (Strong) independence assumption. It was introduced under a different name into the text retrieval community and remains a popular(baseline) method for text categorizing, the problem of judging documents as belonging to one category or the other with word frequencies as the feature. An advantage of Naïve Bayes' is that it only requires a small amount of training data to estimate the parameters necessary for classification. Abstractly, Naïve Bayes' is a conditional probability model. Despite its simplicity and strong assumptions, the naïve Bayes' classifier has been proven to work satisfactorily in many domains. Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined. In Naïve Bayes' technique, the basic idea to find the probabilities of categories given a text document by using the joint probabilities of words and categories. It is based on the assumption of word independence. The starting point is the Bayes' theorem for conditional probability, stating that, for a given data point x and class C:

$$P\,(C\,/\,x) = P(x/C)/P(x) \text{ ------------------- } (1)$$

Furthermore, by making the assumption that for a data point x = {x1,x2,...xj}, the probability of each of its attributes occurring in a given class is independent, we can estimate the probability of x as follows:

$$P(C/x)=P(C).\prod P(xi/C) \text{ --------------------- } (2)$$

5. Visualisation and Analysis

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, neutral sentiment: compound score = 0, negative sentiment: compound score < 0. The positive, neutral and negative scores are ratios for the proportions of data in each tweet.

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

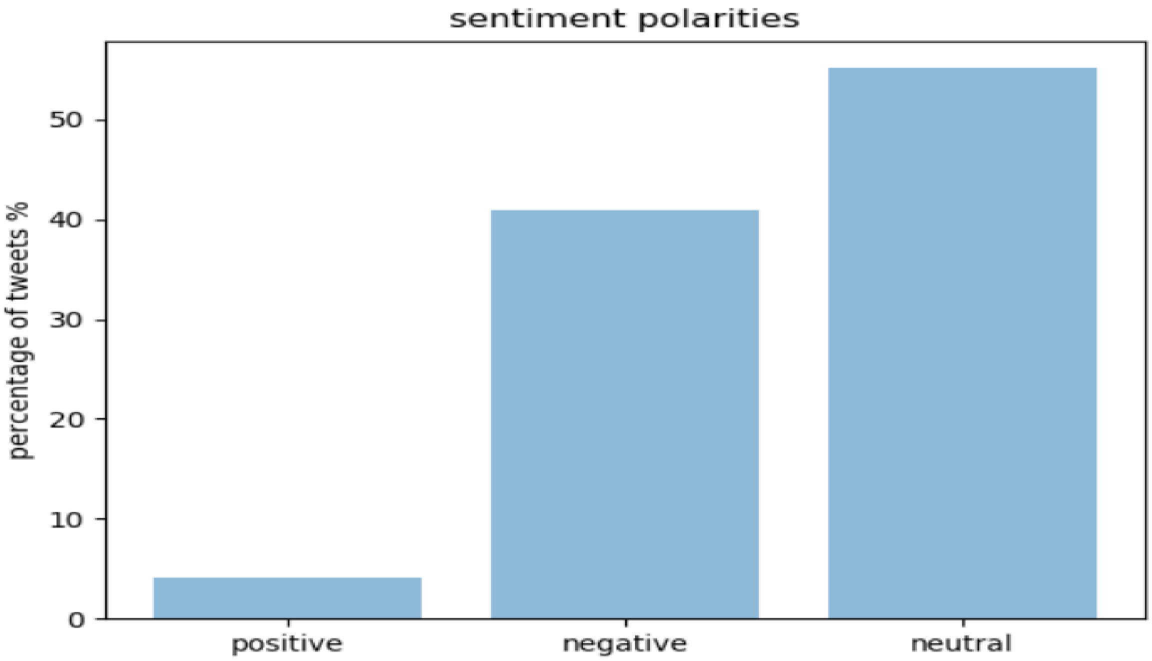

Fig 1. Chart

The compounded scores were divided into three categories as under:

Negative < 0

Neutral =0

Positive > 0

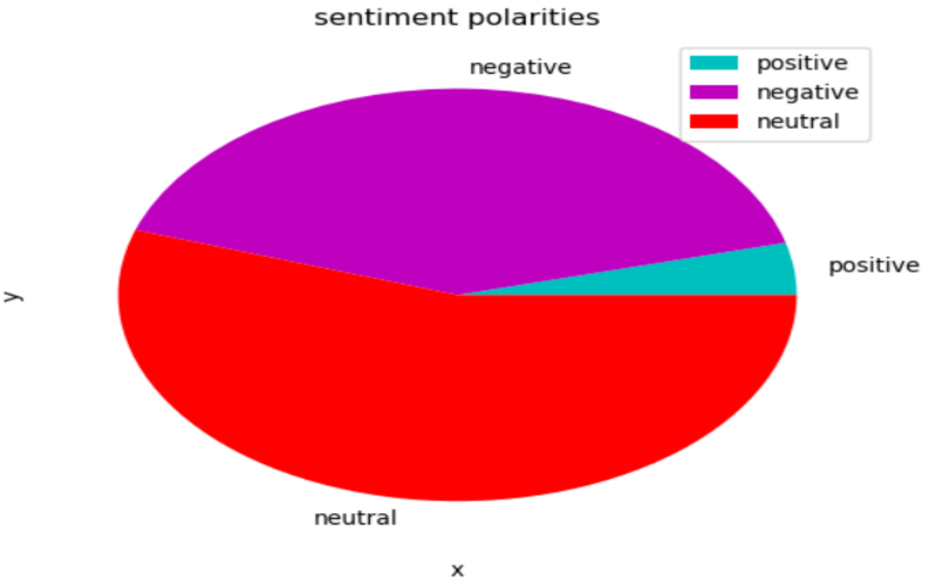

Figure 2 pi chart

6. 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 visualize 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.

References

[1] C. L. Dey and S, "“Canonical PSO Based *K*-Means Clustering Approach for Real Datasets," *International Scholarly Research Notices Hindawi Publishing Corporation*, pp. 1-11, 2014.

[2] R. D. a. S. Chakraborty, "Convex-hull & DBSCAN clustering to predict future weather," in *6th International IEEE Conference and Workshop on Computing and Communication*, Canada.

[3] J. G. Meena Rambocas, "Marketing Research: The Role of Sentiment Analysis," April 2013.

[4] L. W. S. R. Z. Z. Weiguo Fan, "Tapping into the Power of Text Mining," *Journal of ACM*, 2005.

[5] "Movie review dataset," [Online]. Available: http://www.cs.cornell.edu/people/pabo/movie-review-data/.

[6] K. M. Leung, "Naive Bayesian classifier," [Online]. Available: http://www.sharepdf.com/81fb247fa7c54680a94dc0f3a253fd85/naiveBayesianClassifier.pdf.

[7] L. Y. a. X. S. Zhou Yong, "An Improved KNN Text Classification Algorithm Based on Clustering," *Journal of computers*, vol. 4, 2003.

[8] "Twitter dataset," [Online]. Available: http://utils.paranoiaworks.org/diacriticsremover/.