Sentiment Analysis Using Naïve Bayes Algorithm

With Case Study

Jishnusri Ojaswy Akella
Integrated M.Tech, Computer Science
University of Hyderabad
Hyderabad, India
ojaswyajs@gmail.com

LN Yashaswy Akella
Associate Consultant
Capgemini India Pvt. Ltd.
Hyderabad, India
yashaswyaln@gmail.com

***Abstract*— Data is the new oil. It becomes valuable only when mined appropriately. Huge amounts of money is invested in the Data segment in the company only because the companies know its importance. Apart from profit generating information, the companies also need to know the customers' opinions of about the measures implemented. Text mining is the discovery of valuable information from the text in a text file. Online Text mining relates to gathering the information from social media posts such as tweets and blogs. Sentiment Analysis is the computational treatment of opinion, sentiment and subjectivity [4]. Sentiment Analysis of this information extends to the emotions behind this information which assists in selecting appropriate steps. In this paper we chose to mine the opinion of people regarding the two most important measures taken by the Government of India – Demonetization and Goods and Services Tax (GST).**

Keywords—Sentiment Analysis, Machine Learning

I. Introduction

What is a Sentiment?

Sentiments are described as emotions, judgements, beliefs or ideas inspired by emotions. [5]. Opinion Analysis (also known as sentiment mining, sentiment classification, opinion mining, subjectivity analysis, review mining or appraisal extraction, and in some cases polarity classification) deals with using Natural Language Processing, statistics, or machine learning techniques to identify and define the sentiment content of a text unit. [6] Sentiment Analysis and Text mining obtained a great deal of attention since 2010 due to a tremendous amount of data generated online. IDC, in a report sponsored by EMC, predicts that the data volume will grow to 40 zettabytes by 2020, leading to a 50-time growth from the beginning of 2010 [3].

A Little introduction about our case study: Demonetization and GST:

Demonetization is the act of stripping a currency unit of its status as legal tender. It occurs whenever there is a change of national currency: The current form or forms of money is pulled from circulation and retired, often to be replaced by new notes or coins. [1]. Goods and Services Tax (henceforth used as GST in the paper) law in India is an indirect comprehensive tax levied on the supply of Goods and Services. GST replaced a bulk of indirect taxes with a unified tax under different slabs of 0%, 5%, 12%, 18%, and 28% and is set to reshape the country's $2 trillion economy. Demonetization towards the end of 2016 had a huge impact on the lives of people that was stated as a measure to curb illegitimate money. This was followed by the implementation of Goods and Services Tax (GST) law in mid-2017, which gave a question to many economists in the area of national GDP growth. According to Government Data, the result for Demonetization and GST was positive; this paper discussed the opinion from the masses.

Fig.1. Trends of GST over time for a period of 1 year from Dec 2016. Image Source: Google Trends

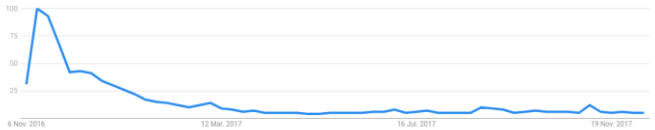

Fig.2. Trends of Demonetization over time for a period of 1 year from November 1st 2016. Image Source: Google Trends

II. Related Work: Classification Techniques

A. *Types of Classification Techniques [9]*

1. Manual Classification

This is an early kind of classification where every piece of information was manually categorized by experts. This is consistent only when the problem size and the team is small. As the information expands, it becomes difficult and expensive to scale. This implies we need automatic classifiers for bigger datasets. This was used by the original (now defunct) Yahoo! Directory [8].

2. Hand-coded rule-based classifiers.

It is generally used by intelligence agencies. It is also widely deployed in Government and other official enterprises. Commercial systems have complex query languages. Accuracy can be higher if a rule is carefully refined over a period of time. It is difficult to build and maintain these rules.

3. Supervised Learning

The majority of practical machine learning uses supervised learning. Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output [10]. It is further classified into two types.

a. Classification: A classification problem is when the output variable is a category like positive or negative and discrete or continuous.

b. Regression: A regression problem is when the output variable is a real continuous value, such as height, temperature.

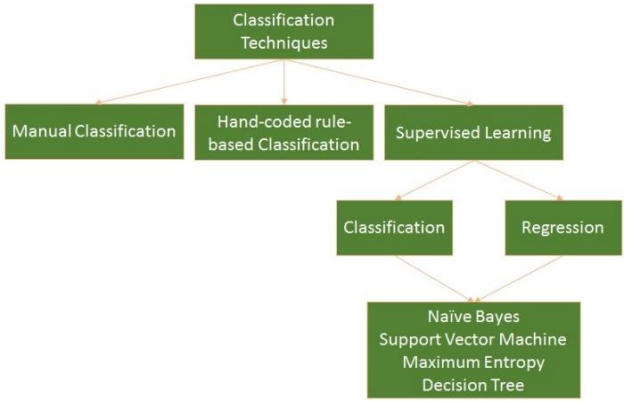

Fig. 3. Classification Techniques and Algorithms.

B. Machine Learning Techniques for Twitter Datasets

Machine learning uses a training set and a test set for classifying the data. Training set contains features and their class labels already recorded by the algorithm. Using this, the system develops its own classification techniques with appropriate labels. When the test cases are given as input, the features of developed training set validate the result.

Different types of algorithms are applied to classify the text. These include Naïve Bayes, Support Vector Machines and Maximum Entropy. The features generally useful for classification of tweets are term frequency, n-grams and parts of speech [11]. Naive Bayes works well for certain problems with highly dependent features. This is unique given that the features in Naive Bayes are independent [12].

A new model was introduced in which efficient approaches are used for feature selection, weight computation and classification. The model is built on Bayesian algorithm [13]. One research proposed an influential probability model for twitter sentiment analysis [14]. If @text is found in the body of a tweet, it is influencing action and it contributes to influencing probability. Any tweet that begins with @text is considered as a retweet (shared from original tweet). It implies an influenced action from the original tweet and it contributes to an enhanced influenced probability. Another research included a 2-step automatic sentiment analysis method for classification of the tweets. A noisy training set was introduced to reduce the labeling effort while developing classifiers. Tweets were classified into subjective and objective tweets. Later, the subjective tweets are classified as positive and negative tweets [15].

III. METHODOLOGY

Naïve Bayes was chosen as the classifier for this project as it is easy and fast to predict the class of the test data set. It also performs well in the multi class prediction. It perform well in case of categorical input variables compared to numerical variable(s) [16]. Since tweets are classified into different categories of information, Naïve Bayes was chosen.

Naïve Bayes working is not discussed in the paper, but referred as it is beyond scope. We chose this algorithm owing to the accuracy it provides. Twitter was mined for the period of 5 months after the GST was launched on 1st July 2017. In this period 10000 tweets were collected and analysis was performed. This included the following steps in the project

1. Collection of Data

Data was collected for 150 days. This included a random selection of tweets with the hashtags GST and Demonetization.

2. Pre Processing the Twitter Data

The Data from Twitter is to be processed clearly. It has a lot of misspellings owing to the character limit it gives. The information could be clear but due to such errors the, the algorithm will be affected. Some tweets included the URLs and also some slang words which influence the emotion of the tweet. Hence we made sure that these URLs were removed and slang words were classified according to their intensity and relevance before taking a stand on the tweet. Hence we cleaned all the 10000 tweets before beginning the processing of our data. Some of the tweets were removed because they contained the same hashtag but did not contain any information.

3. Setting up the python workspace.

Anaconda suite was used and Naïve Bayes algorithm was coded in python 3.

IV. RESULTS

There are two major entities taken into consideration. They are Subjectivity and Polarity. A sentence that states a fact is termed Objective whereas a sentence expressing an opinion is referred as subjective. Here we give a Full Subjective Statement 1 and a full Objective Statement -1. Polarity is the emotions expressed in a sentence. In this research, +1 is a highly polar statement and -1 is a strong disapproval on a topic. Based on this, the results are presented below.

I. GST Results

The following are the results of polarity of Latest tweets, popular tweets and randomly selected tweets for GST.

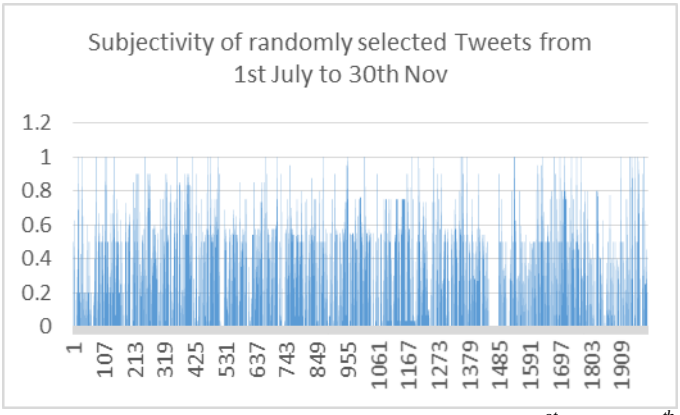

Fig. 4. Subjectivity of GST (random tweets) from 1st July to 30th November 2017.

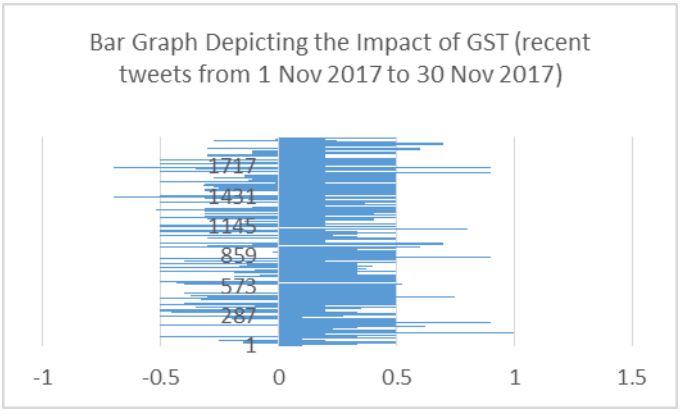

Fig. 5. Bar Graph Depicting the Polarity Impact of GST (recent tweets from 1 Nov 2017 to 30 Nov 2017)

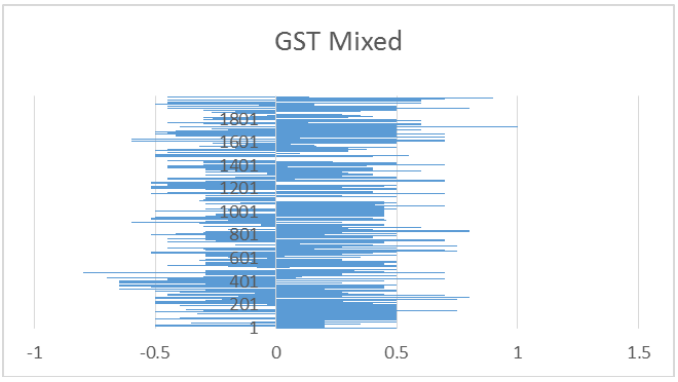

Fig. 6. Polarity of a set of Tweets randomly selected from 1st July to 30th Nov 2017

Understanding the GST Results: All the tweets from GST are subjective: i.e. all of them are opinions of people. When it was introduced on 1st July, due to misinformation, it was understood to be a destroyer of the economy, but as time progressed and as the GST slab rates fell down (only 50 items currently are in 28% tax slab[16]), people saw a positive change. Hence, there was a transition from negative to positive**.

II. Demonetization Results

The following are the results of polarity of Latest tweets and randomly selected tweets for Demonetization.

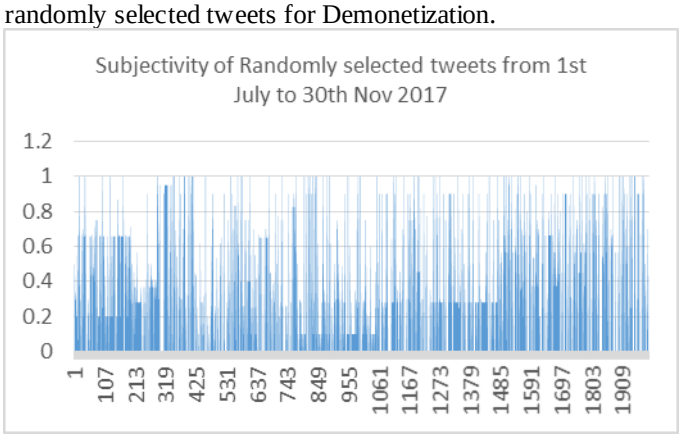

Fig. 7. Subjectivity of Demonetization (random tweets) from 1st July to 30th November 2017.

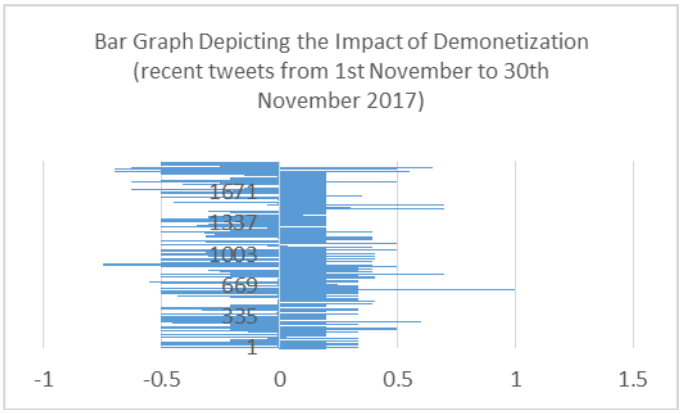

Fig. 8. Bar Graph Depicting the Polarity Impact of Demonetization (recent tweets from 1 Nov 2 to 30 Nov 2017)

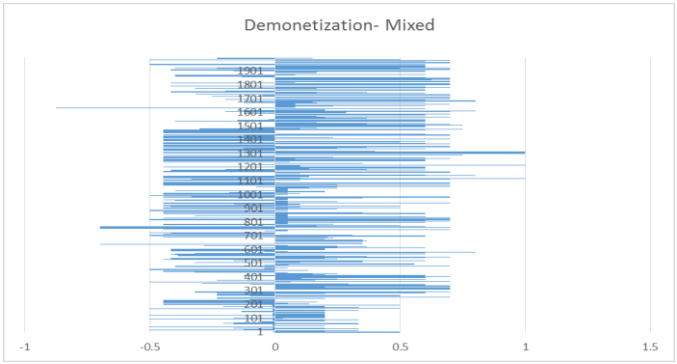

Fig. 9. Polarity of a set of Tweets randomly selected from 1st July to 30th Nov 2017

Understanding the Demonetization Results: All the tweets from Demonetization are subjective: i.e. all of them are opinions of people. The recent tweets in the month of November show a moderate disapproval from the people. It was also the first anniversary of Demonetization. However, when the data is taken over the period of 5 months, people were positive on the implementation of Demonetization**.

V. CONCLUSION

Sentiment analysis can be performed using Manual Classification (Lexicon Based) approach or machine learning or an implementation of both (hybrid) approach. The lexicon based approach is at a disadvantage when the size of the data is huge and the availability of experts is limited. Hence, Machine Learning Algorithms are preferred. This report deals with Naïve Bayes algorithm and its execution to a trending issue in the country, Demonetization and Goods and Services Tax. Since many of the users share their opinion on social media websites such as twitter, this was taken to collect the information. Statistics and Visualization techniques are implemented on the Polarity and Subjectivity of the Tweets. A full-fledged end-to-end python code was written to mine the data and later classify to derive the solutions.

ACKNOWLEDGMENT

We would like to thank our mother, Dr. D Lalitha Devi, Head, Department of Statistics, K.G.C.W. for her support in this project. We would like to thank our father, Shri A.S.N. Murthy for his encouragement throughout our project.

REFERENCES

[1] Definition of Demonetization, Investopedia, "https://www.investopedia.com/terms/d/demonetization.asp"

[2] Mehdi et all, A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques

[3] John Gantz and David Reinsel. 2012. THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Grow th in the Far East. Technical Report 1. IDC, 5 Speen Street, Framingham, MA 01701 USA. Accessed online on May, 2017. https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf

[4] Sarah Schrauwen, Machine Learning approaches to sentiment analysis using the Dutch Netlog Corpus, CLiPS Research Center, University of Antwerp, 2010.

[5] BOIY, E., HENS, P., DESCHACHT, K. & MOENS, M.-F. (2007), "Automatic Sentiment Analysis in OnLine Text". In Proceedings of the Conference on Electronic Publishing (ELPUB-2007), p. 349-360

[6] MEJOVA, Y. (2009), "Sentiment Analysis: An Overview". Comprehensive exam paper, available on http://www.cs.uiowa.edu/~ymejova/publications/CompsYelenaMejova.pdf [2010-02-03].

[7] https://en.wikipedia.org/wiki/Yahoo!_Directory

[8] Introduction to Information Retrieval, Stanford on Coursera.

[9] https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/

[10] Y. Mejova, "Sentiment analysis: An overview," Comprehensive exam paper, available on http://www. cs. uiowa. edu/~ ymejova/publications/CompsYelenaMejova.pdf [2010-02-03], 2009.

[11] P. Domingos and M. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," Machine Learning, vol. 29, no. 2-3, pp. 103–130, 1997.

[12] Z. Niu, Z. Yin, and X. Kong, "Sentiment classification for microblog by machine learning," in Computational and Information Sciences (ICCIS), 2012 Fourth International Conference on, pp. 286–289, IEEE, 2012. [12] L. Barbosa and J. Feng, "Robust sentiment detection on twitter from biased and noisy data," in Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 36–44, Association for Computational Linguistics, 2010.

[13] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in Proceedings of LREC, vol. 2010, 2010.

[14] A.Celikyilmaz,D.Hakkani-Tur,andJ.Feng,"Probabilisticmodel-based sentiment analysis of twitter messages," in Spoken Language Technology Workshop (SLT), 2010 IEEE, pp. 79–84, IEEE, 2010.

[15] "What are the Pros and Cons of Naive Bayes?", https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/, 2017

[16] Only 50 items at 28% GST from The Hindu Business Line http://www.thehindubusinessline.com/economy/policy/gst-council-prunes-list-of-goods-to-be-taxed-at-28-to-just-50/article9952407.ece.

**** This data is taken from twitter and information is inferred. The authors hold no responsibility of the views generated from this research paper.**

Disclaimer: The Data in this paper and the authors are not endorsing any kind of impact of Demonetization and GST. We presented the views from the tweets mined (opinions of the people). We hold no responsibility to the information presented and the views inferred from the Data.

ABOUT THE AUTHORS

Jishnusri Ojaswy Akella is an undergraduate student of Computer Science pursuing his Integrated Master of Technology degree from University of Hyderabad. He is interested in the areas of Machine Learning and Artificial Intelligence.

LN Yashaswy Akella received his B.E. (Hons.) Degree in Electrical and Electronics Engineering from BITS Pilani. He is currently an Associate Consultant with Capgemini. His research interests include Business Analytics, Statistics and Data Sciences. He has previous written 3 research papers in the same areas.