# A Data Mining System for Sentiment Classification of Indian Currency Demonetisation using Naive Bayes Classifier Algorithm

**Conference Paper** · March 2017

CITATIONS
0

READS
511

**1 author:**

Athul Jayaram
Amity University
**12** PUBLICATIONS **10** CITATIONS

SEE PROFILE

# A Data Mining System for Sentiment Classification of Indian Currency Demonetisation using Naive Bayes Classifier Algorithm

Athul Jayaram

Amity University Uttar Pradesh

athuljayaram@gmail.com

+91-8285188306

## Abstract

Demonetisation is the retirement of the existing currency with the new currency. Demonetisation aims to control the flow of black money, prevent corruption, curb financing of terrorism and most importantly form a cashless economy that is transactions is conducted through online. During demonetisation, people used debit cards, credit cards, mobile wallet and net banking for their day today need. Demonetisation lead to scarcity of cash as well as low spending and fewer purchases by the consumers. Twitter tweets of 11843 people were collected and data mining was conducted to understand the real sentiment of the people of India during the demonetisation period using naive Bayes classifier algorithm. T test and F test were conducted to determine the sentiment of the people using hypothesis. The test results of the study were very convincing and strong that the people had a negative sentiment towards the Indian currency demonetisation.

## Keywords

Data Mining, Sentiment Classification, demonetisation, Naïve Bayes, Classification Algorithm

## Introduction

Demonetisation is the replacement of old currency with the new currency. The existing currency of a country is retired and the new currency is introduced. Often the new currency is different from the existing currency and government introduces various measures for citizens such as by exchanging the old currency by depositing at the bank and obtain the new currency from the bank. Demonetisation is often conducted by the government with different objectives. In India, the demonetisation that took place in November 2016 was mainly to control the fake currency, tackle black money, corruption, curb financing of terrorism and form the cashless economy.

The government demonetized currency of value of rupees 500 and 2000 which accounted for over 86 percentage of the total cash supply in India. All individuals processing currency of value 500 and 2000 where to deposit same to the bank and show the proof that the taxes of the same amount have been paid already else tax penalty of 200% is charged on the deposited money. Demonetisation led to the scarcity of cash and lower spending by the consumers. Consumer started using various cashless payment systems such as card payments, online bank transfers, and digital payment mobile applications.

## Literature Review

Various researches have been conducted about demonetisation. In Gajjar, (2016) the present and future challenges of demonetisation have been discussed. The darker side of the economy and the black money is discussed on (Jain, 2016 ) (Chandel, 2017), while demonetisation and its impact on the society are discussed on (Ganiger & Ranganatha, 2017). The success of digital

transactions is discussed on (Mehta, Patel, Mehta, & others, 2016). The factors that led to the shortage of cash and the black money is discussed on (Lahiri & others, 2016), while paper of (Rani s., 2016) portrait demonetisation as a tool used by the prime minister to root out black money and terrorist activities from indian economy. While (Pachare, 2016) gives a complete overview about the digital wallets in India. (Mahajan & Singla, 2017) (Singh & Singh, 2016) discusses the financial stability of India during the time of demonetisation. (Rani g., 2016) discusses the effect of demonetisation on the retail outlets. While (Krishnan & Siegel, 2017) has conducted a good research about the impact of demonetisation and around 28 slums of Mumbai city of India. (Sharma, 2016) gives a good estimation of the size of black money in India. (Green, 2016) (Rakesh, 2017) gives a critical analysis of Indian currency demonetisation. (Gupta, 2017) discusses different banking channels for cashless growth. (Jain M. , 2017) discusses challenges and opportunities for making India a cashless economy. (Mukhopadhyay, 2016) (Rajendran & Maggo, 2017) gives a deep understanding about the cashless payments and impact of it on the economy of India. A detailed explanation of parallel economy is given in (Sarkar, 2012)

## Data Mining

Data mining is the conversion of raw data to useful information (Bramer, 2007). Various data mining software to identify relationships and patterns in the data which can be used by the enterprises to understand the customers and increase sales and revenue (Bharati & Ramageri, 2010).

Data mining involves the following stages such as selection, pre-processing, transformation, data mining and interpretation (Han, Pei, & Kamber, 2011). Selection refers to acquiring of data which is relevant to the area of the problem in which the research is conducted (Hand, Mannila, & Smyth, 2001). It may involve raw data set which has to be for the process through

the pre-processing. Various methods are deployed to clean the data that remove any noisy data, missing data or any other general errors which can occur in the data. After preprocessing the data can be transformed using any transformation process. In transformation process, data is converted from source format to the destination format. Here the destination format is the data format accepted by the data mining system. Once data transformation is carried out well the actual data mining can be done. From data mining, interpretations and conclusions can be drawn out.

Commonly used data mining method are a regression, classification, clustering, association and anomaly detection (Olson & Delen, 2008). Regression creates a function that can create a model of all the data in the data set. Classification is the identification of data and categorises them into specific types. Clustering is the grouping of data into a unique class which has same characteristics. Association rule finds a relationship between the different variables of data, for example, rainfall variable may be related to temperature variable and this can be proved from the data collected and by identifying the relationships between data.

## Naive Bayes Classifier

Naive Bayes classifier is a type of probabilistic classification algorithm that uses Bayes Theorem to classify the data. It is supplied by having a list of words that are positive and also that are negative. The algorithm checks each and every word in every sentence and identifies whether it is a positive word or a negative word. If it is a positive word then a positive one score is given to the word, if it is a negative word, then a negative one score is given to the word. Similarly, the score for every word in a sentence is calculated and the net score is calculated by adding up all the individual value of each word in a sentence. It doesn't take into consideration the order of words nor the grammar of the sentence. It can be used for sentiment

analysis, email spam detection, document categorization and in many other fields (Bifet & Frank, 2010).

The conditional probability by Bayes Theorem can be calculated as

$$P(A \mid B) = P(B \mid A) \times P(A) \div P(B)$$

Where A and B are events and P(B) not equal to 0

P(A) is probability of A while P(B) is probability of B

P(A | B) is the probability of observing event A when event B is true while P(B | A) is the probability of observing event B when event A is true.

## Tweets Sentiment Analysis

The sentiment of the Twitter tweets can be analyzed using naive Bayes algorithm and Bayes theorem (Agarwal, Xie, Vovsha, Rambow, & Passonneau, 2011). Every tweet is taken and each word is matched with the database of positive and negative words (Jiang, Yu, Zhou, Liu, & Zhao, 2011) (Pak & Paroubek, 2010). The probabilities are calculated using Bayes Theorem as follows.

$$P(\text{positive} \mid \text{tweet}) = P(\text{tweet} \mid \text{positive}) \times P(\text{positive}) \div P(\text{tweet})$$

Where positive are the tweets which are positive and tweet is the complete dataset of tweets P(positive) is probability of positive tweets while P(tweet) is probability of whole tweets dataset

P(tweet | positive) is the probability of tweet when a tweet is positive and is true while P(positive | tweet) is the probability of tweet being positive when the dataset is tweeted and is true.

## Dataset

Demonetisation data is collected from Twitter from November 8, 2016, to January 27, 2017. Though there were limitations of the access to the Twitter API, 11843 tweets were collected successfully. The collected data were further processed. Data such as numbers, symbols, non-english characters and extra spaces were removed so that the data is perfect to perform data mining. The collected data is stored in the database of R and various functions are utilized to conduct sentiment analysis (Gokulakrishnan, Priyanthan, Ragavan, Prasath, & Perera, 2012).

## Approach

For better accuracy, sentiment analysis is performed by two methods such as lexicon analysis and naive Bayes classifier. Lexicon analysis includes a words list which is positive and negative. Each tweet is parsed and matched with the words from Lexicon list. Based on the count of positive and negative words, a net score is given for all tweets individually.

The tweets online based on polarity and emotions. Different polarities are positive, negative and neutral. Different emotions are anger, disgust, fear, joy, sadness, surprise and other unknown emotions. The first step is that sentiment analysis of tweets based on polarity is calculated and a graph is plotted. The second step is that sentiment analysis of tweets based on emotions is calculated and further the graph is plotted. In the next step, the words that are frequently used in the tweets are collected and a word cloud is generated for it.

For accurate analysis and to perform data mining, a histogram of demonetisation using lexicon analysis is plotted. Similarly, another histogram of demonetisation by Naive Bayes is also plotted. Further a combined histogram was plotted to compare the accuracy. Curves were also plotted and their bandwidths were also calculated.

## Classified Data Samples

Samples of positive tweets

"and I must mention that response from citizens to decision of demonetisation is positive"

"must watch video beautifully depicts the positive impact of demonetisation commendable effort by Ji"

“with demonetisation, the use of digital payments well have a better cleaner economy”

"remonetizing India will be a better option now coz the effects of demonetisation is crystal clear"

“the biggest thumbs up to demonetisation is the fact that credit default risk of SBI hits yr low”

Samples of negative tweets

"demonetisation impact property sales to see fall, demonetisation disaster"

"India from farm to loom the textile industry suffers after demonetisation"

"demonetisation hit small and medium-sized enterprise sectors output wages reports"

"hundreds of people died standing in queue of ATM over demonetisation”

"demonetisation mess senior citizens lock up bank for denial of pension forced to get change to get pension"

## Study of Data from Reserve Bank of India

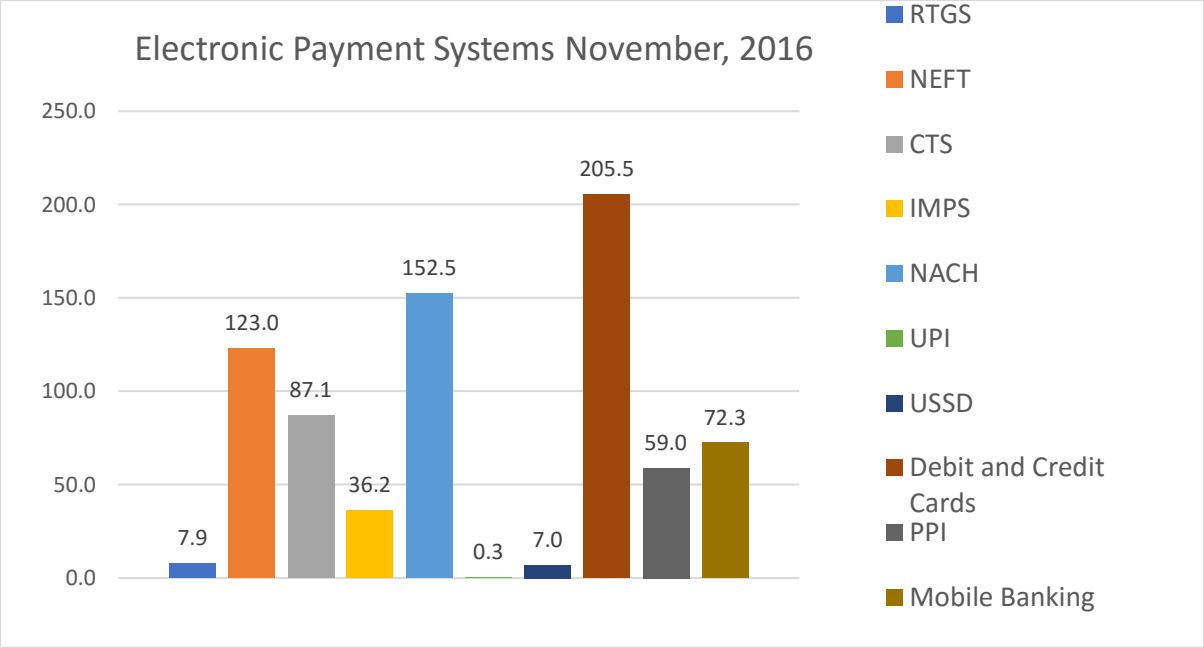


Fig. No. 1 – Electronic Payment Systems Bar Graph, November 2016

As in the Figure No. 1, The electronic payment systems transactions of November 2016 is shown. On November 2016, transactions using debit and credit cards, NACH and NEFT are most used by the consumers. 205.5 million Debit and Credit Card transactions occurred the most in November 2016. About 152.5 million transactions occurred of NACH, 123 Million transactions were of NEFT, 87.1 million transactions where is done by CTS, 72.3 million transactions were done by mobile banking, 59 million transactions were conducted by PPI, 36.2 million transactions were conducted by IMPS, 7.9 million transactions were done by RTGS. Only around 7000 USSD transactions took place in November 2016. Approximately 0.3 million transactions where UPI as in November 2016.

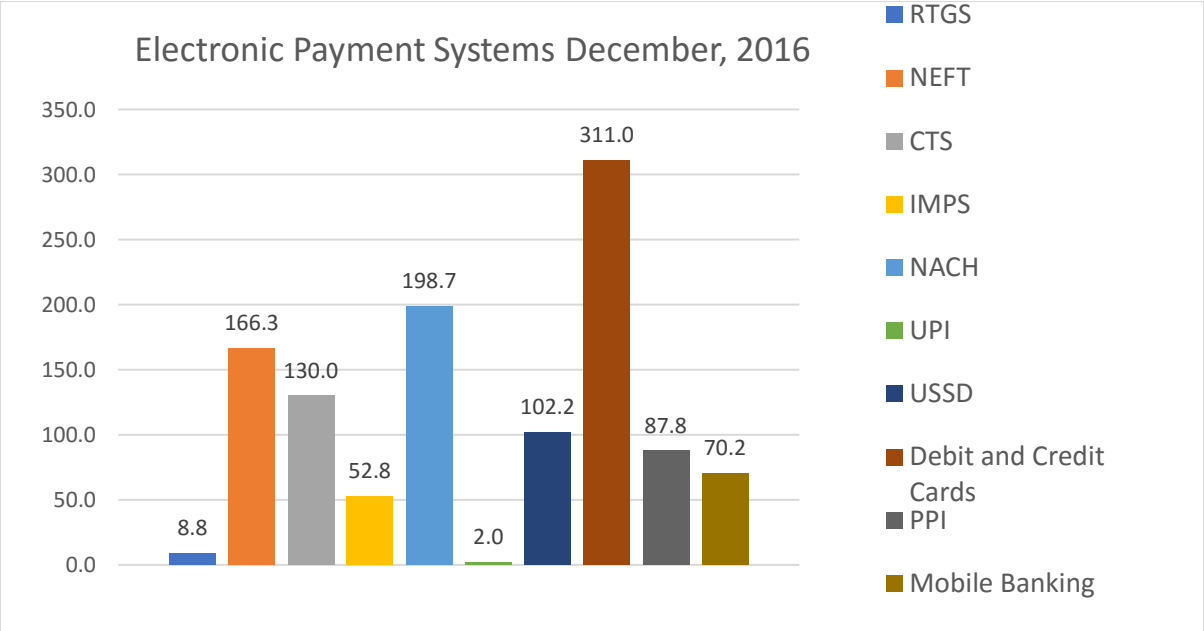


Fig. No. 2 – Electronic Payment Systems Bar Graph, December 2016

In Figure No. 2, the electronic payment systems transactions of December 2016 are shown. On December 2016, transactions using debit and credit cards, NACH and NEFT are most used by the consumers. The debit and credit card transactions amounted to 311 million which is the largest form of electronic payment system being utilized in December. About 198.7 transactions were performed through NACH, 166.3 million transactions were done using NEFT, 130 million CTS, 102.2 thousand USSD transactions, 87.8 million PPI transactions, 70. 2 million mobile banking transactions, 52.8 million IMPS transactions, 8.8 million RTGS transactions, the least is UPI transactions which are only around 2 million transactions.

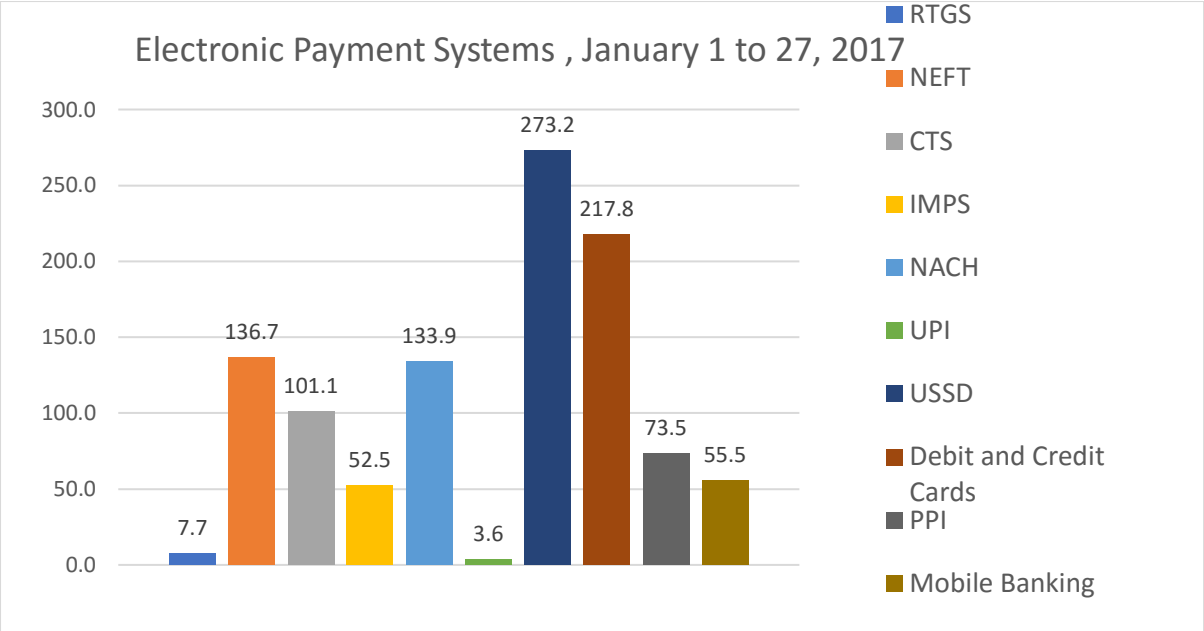


Fig. No. 3 – Electronic Payment Systems Bar Graph, January 1, 2017, to January 27, 2017

Electronic payment systems of January 2017 shown in the Figure No. 3. On January 2017, transactions using USSD, debit and credit cards and NEFT are most used by the consumers. 273.2 thousand USSD transactions took place, while 217.8 million transactions were done using debit and credit cards, about 136.7 million transactions were done using NEFT, about 133.9 million transactions using NACH, about 101.1 million transactions using CTS, 73.5 million transactions using PPI, 52.5 million transactions using IMPS, 7.7 million transactions using RTGS, 3.6 million transactions using UPI.

Table No. 1 – Top 10 Banks Operating in India by the no. of Transactions in December 2016

| Name of Bank | No. Of Transactions (December, 2016) |
|---|---|
| STATE BANK OF INDIA | 55598349 |
| HDFC BANK | 34713458 |
| ICICI BANK LTD | 24906592 |
| AXIS BANK | 23314280 |

| | |
|---|---|
| PUNJAB NATIONAL BANK | 12783229 |
| BANK OF BARODA | 10753510 |
| BANK OF INDIA | 9044864 |
| YES BANK | 8586957 |
| CANARA BANK | 8010186 |
| IDBI BANK | 7892838 |

The top 10 banks based on the number of transactions is given in the Table No. 1. State Bank of India, HDFC Bank, ICICI Bank and Axis Bank tops the list. State Bank of India customers has performed 5,55,98,349 transactions , HDFC Bank has performed 3,47,13,458 transactions, ICICI Bank has performed 2,49,06,592 transactions and Axis Bank has performed 2,33,14,280 transactions, while the lowest is IDBI bank with 78,92,838 transactions. During demonetisation people has preferred the banks which are top on the list for their purpose and can be considered as banks which were trusted by people.

Results

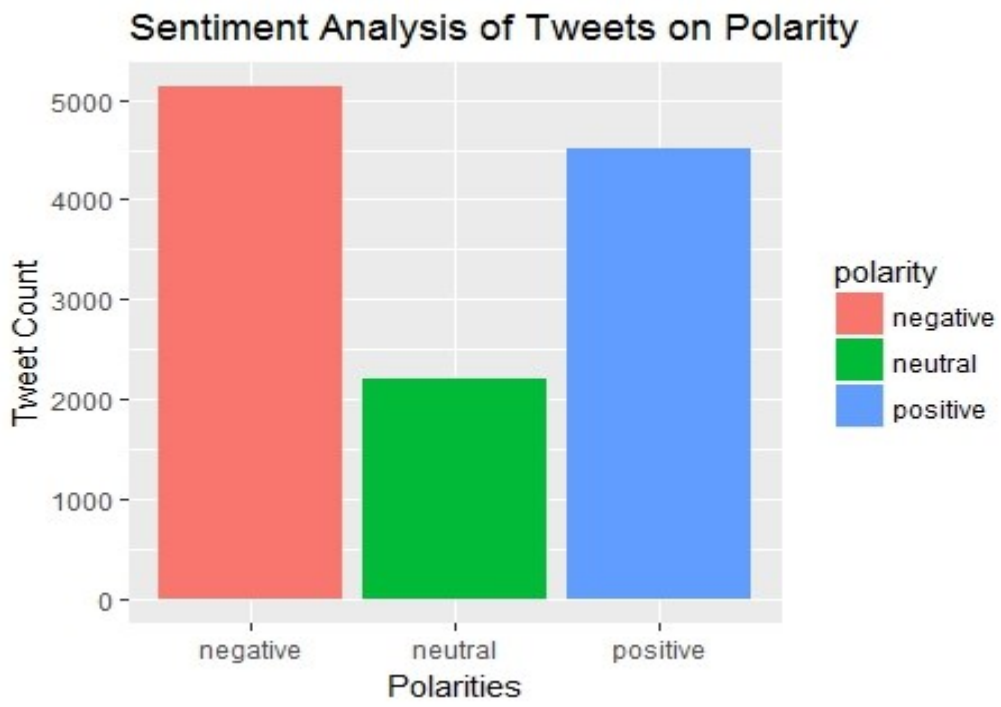


Fig. No. 4 – Sentiment analysis of Tweets on Polarity

Sentiment analysis was performed on 11843 tweets as shown in Figure no. 4. The polarity was identified with over 5200 tweets with negative polarity, around 2000 neutral polarity tweets 4500 positive tweets.

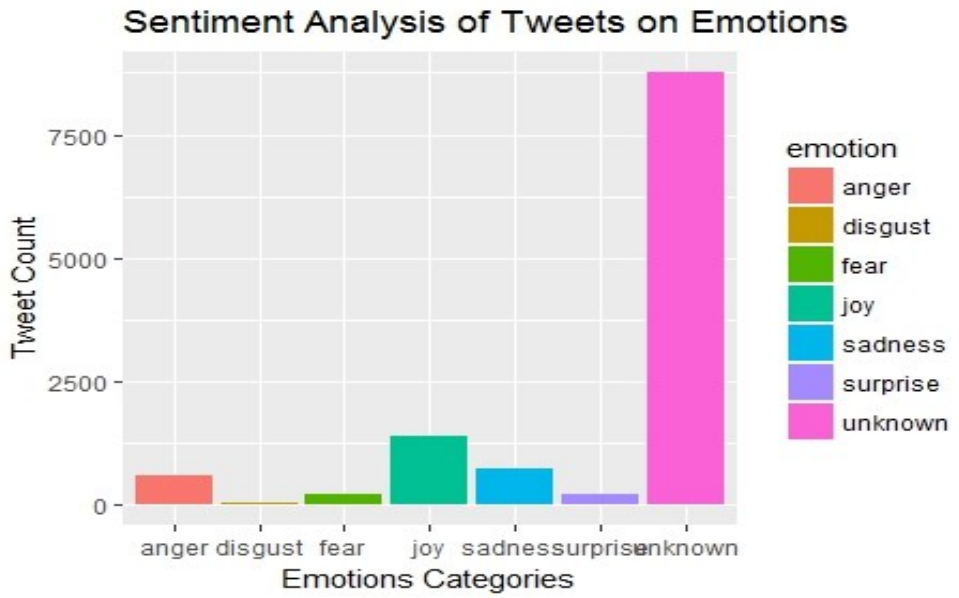

Fig. No. 5 – Sentiment analysis of Tweets on Emotions

Sentiment analysis of tweets based on emotions was performed and the result was plotted as shown in Figure no. 5. The different emotions such as anger, disgust, fear, joy, sadness, the surprise was identified. Emotions such as joy, sadness, and anger were the emotion of the most of the people during demonetisation of currency in India.

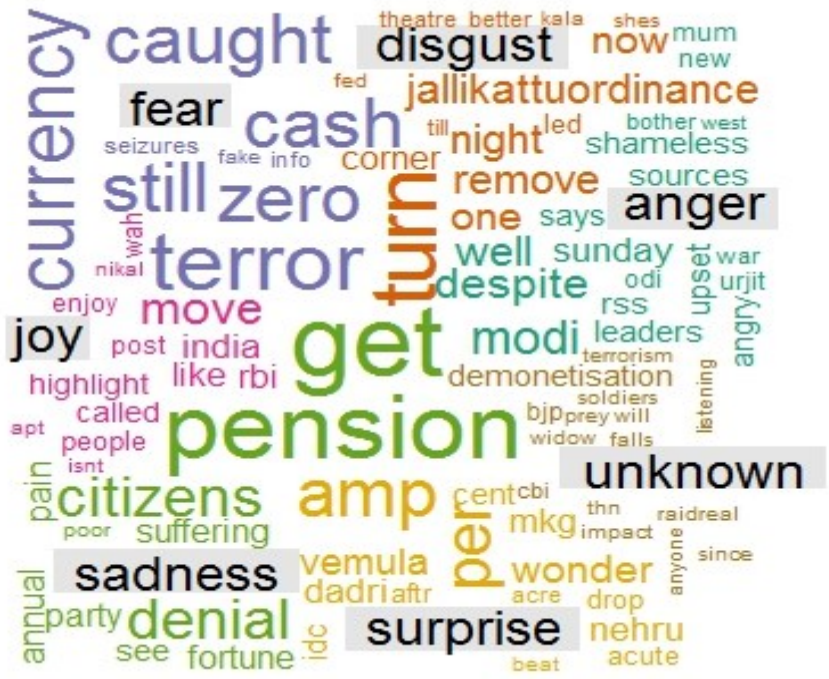


Fig. No. 6 – Twitter Tweets Cloud

A word cloud was formed for the most repeated words used in tweets by everyone who was conversing about demonetisation as shown in Figure No. 6.

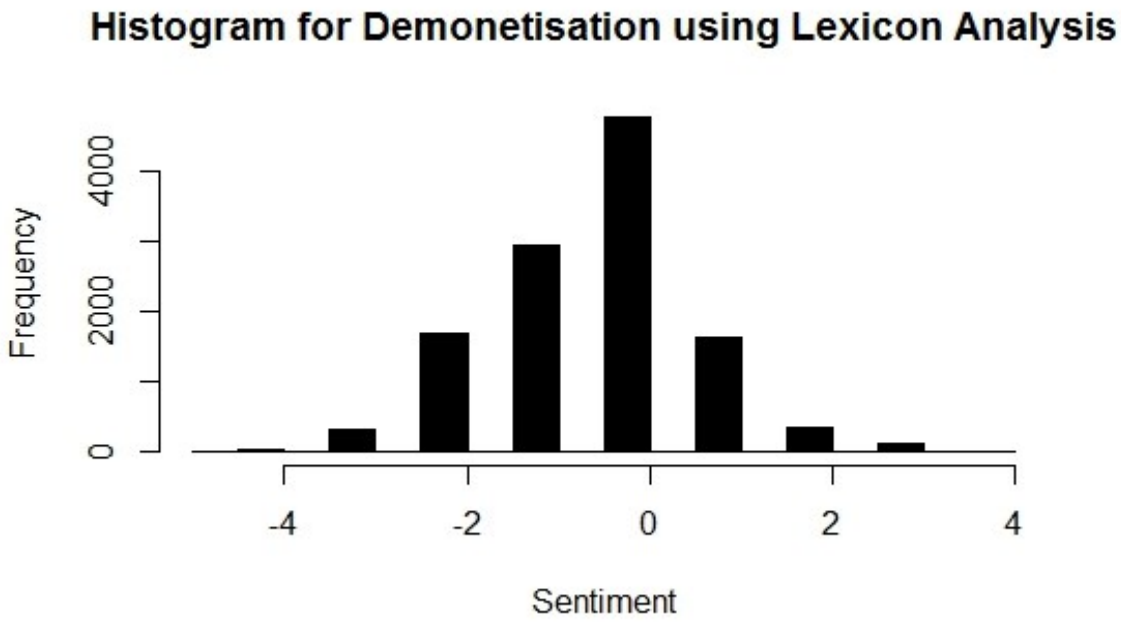


Fig. No. 7 – Histogram for demonetisation using Lexicon Analysis

Histogram for demonetisation using Lexicon Analysis is plotted with the sentiment on the x-axis and frequency on the y-axis as shown in Figure. No. 7. Clearly, it shows the sentiment of the majority of people is below zero or negative.

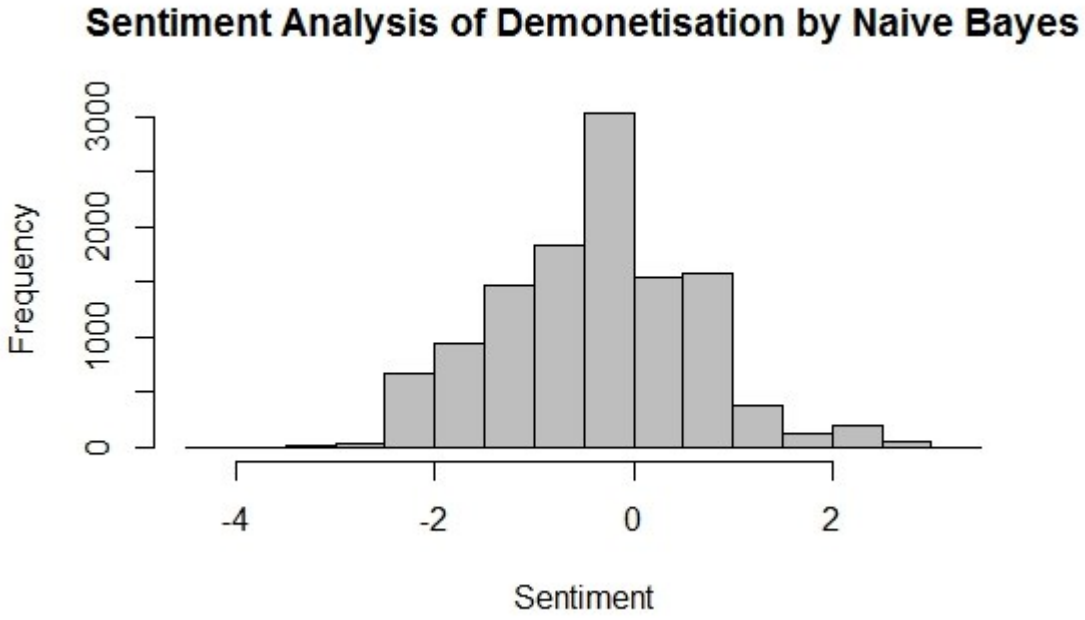

Fig. No. 8 – Sentiment Analysis of demonetisation using Naive Bayes Classifier

Histogram for demonetisation using Naive Bayes Classifier is plotted with the sentiment on the x-axis and frequency on the y-axis as shown in Figure. No. 8. It shows the sentiment of the majority of the people is below zero or negative.

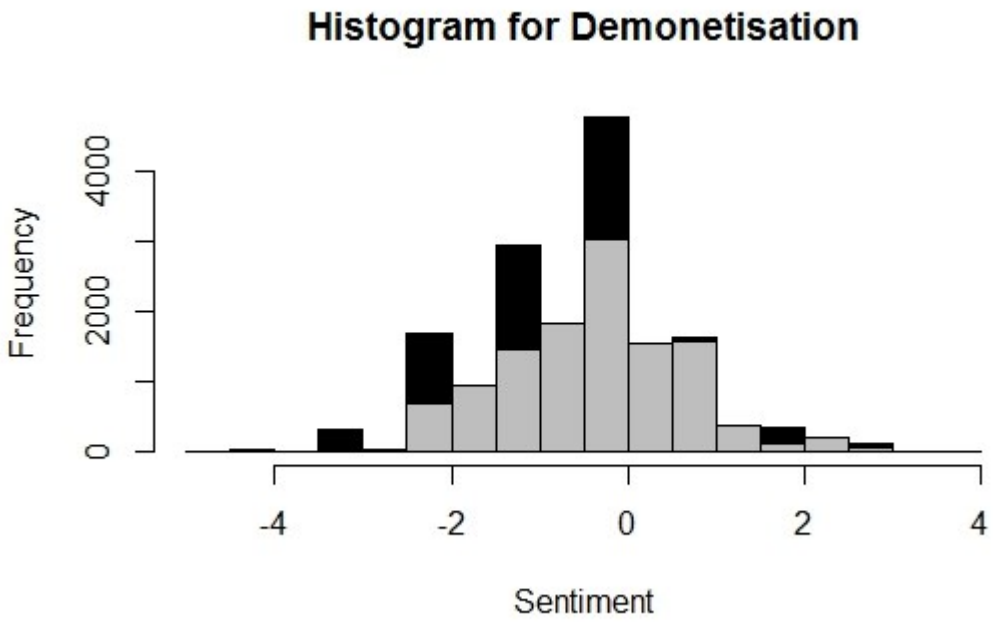


Fig. No. 9 – Histogram for demonetisation by Naive Bayes Classifier and Lexicon Analysis

A combined histogram for demonetisation using Lexicon Analysis and Naive Bayes Classifier is plotted with the sentiment on the x-axis and frequency on the y-axis as shown in Figure No. 9. Here black color indicates lexicon analysis while gray indicates Naive Bayes Classifier. It shows the sentiment of the majority of the people is below zero or in the negative range.

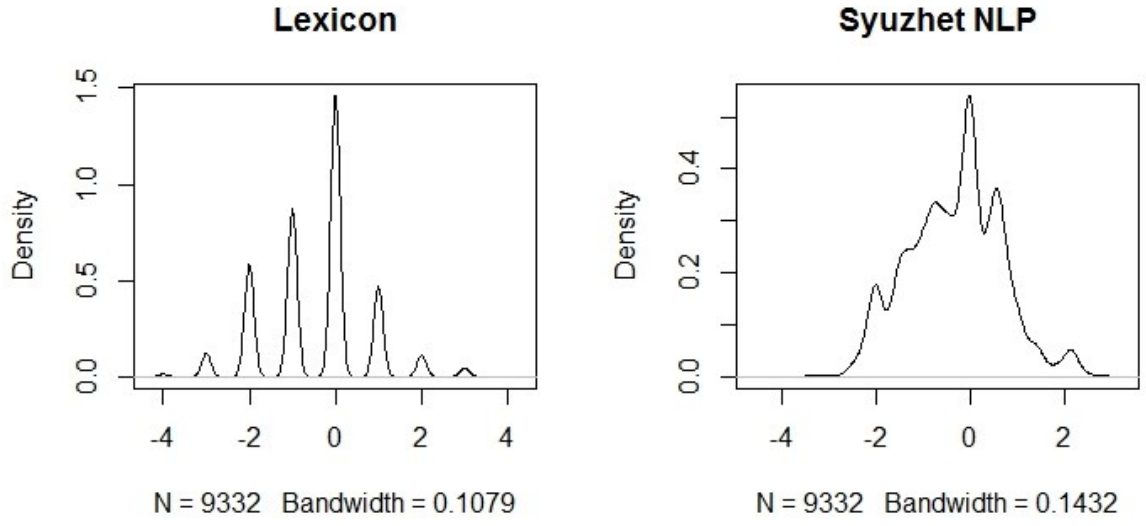


Fig. No. 10 – Lexicon analysis and Syuzhet NLP curves

Curves of Lexicon analysis and Syuzhet NLP were plotted as shown in Figure No. 10. Lexicon has a bandwidth of 0.1079 while Syuzhet NLP used for Naïve Bayes Classifier has a bandwidth of 0.1432.

Table No. 2 - Syuzhet NLP and Lexicon Analysis Results

| | **Syuzhet NLP** | **Lexicon** |
|---|---|---|
| **Population Mean** | -0.438 | -0.313 |
| **Standard Deviation** | 0.990 | 1.181 |
| **Median** | -0.25 | 0 |
| **Minimum Sample Value** | -4.2 | -4 |
| **Maximum Sample Value** | 2.85 | 4 |
| **Confidence Interval of T-test** | -0.165 | -0.083 |
| **Confidence Interval of F-test** | 1.481 | 1.365 |
| **A Confidence Level** | 95% | 95% |
| **One Sample One-Tailed T-test** | -30.624 | -35.863 |
| **T-test value when α = 0.05** | 1.645 | |
| **One-tailed F-test** | 1.422 | |
| **F-test value when α = 0.05** | 1.645 | |

## One-tailed one sample t-test of Syuzhet NLP analysis

The analysis result is shown in the Table No. 2. The mean sentiment score of the sample (11843 tweets) is -0.438. The one-sided 95% confidence interval tells us that mean sentiment score of the population is likely to be less than -0.165. Since the p-value is less than the significance level of 0.05, we have strong evidence to reject the null hypothesis that the mean sentiment score is equal to zero. This means that there is strong evidence that the population's sentiment on the demonetisation is Negative (Alternative hypothesis).

## One-tailed one sample t-test of Lexicon analysis

The mean sentiment score of the sample (11843 tweets) is -0.313. The one-sided 95% confidence interval tells us that mean sentiment score of the population is likely to be less than -0.083. Since the p-value is less than the significance level of 0.05, we have strong evidence to reject the null hypothesis that the mean sentiment score is equal to zero. This means that there is strong evidence that the population's sentiment on the demonetisation is Negative (Alternative hypothesis).

## Conclusion

Twitter Sentiment analysis to understand the people’s sentiment on demonetisation put forward by the government. Methods such as Naïve Bayes and Lexicon are used for analyzing the sentiment of the tweets. T-test and F-test are conducted to determine the sentiment of the people, further, the hypothesis testing is performed. The Hypothesis test results gave strong evidence that the people of India are having negative sentiment towards the Indian Currency demonetisation.

## References


1. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of twitter data. *Proceedings of the workshop on languages in social media*, (pp. 30-38).
2. Bharati, M., & Ramageri, M. (2010). Data mining techniques and applications.
3. Bifet, A., & Frank, E. (2010). Sentiment knowledge discovery in twitter streaming data. *International Conference on Discovery Science*, (pp. 1-15).
4. Bramer, M. (2007). *Principles of data mining* (Vol. 180). Springer.
5. Chandel, A. S. (2017). Demonetisation in India: An Overview. *ICPESS 2017-Sarajevo BOSNIA*.
6. Gajjar, N. (2016). Black Money in India: Present Status and Future Challenges and demonetisation. *International Journal, 4*.
7. Ganiger, S. V., & Ranganatha, B. (2017). Demonetisation and its impact on social development. *Indian Journal of Applied Research, 7*.
8. Gokulakrishnan, B., Priyanthan, P., Ragavan, T., Prasath, N., & Perera, A. (2012). Opinion mining and sentiment analysis on a twitter data stream. *Advances in ICT for emerging regions (icter), 2012 International Conference on*, (pp. 182-188).
9. Green, R. A. (2016). Will India's Currency Exchange Work as Hoped? *Issue Brief*.
10. Gupta, D. K. (2017). Demonetisation in India 2016--Mother Tongue Friendly E-Delivery Banking Channels for Cashless Growth.
11. Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques.* Elsevier.
12. Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining.* MIT press.

13. Jain, L. (2016). Black Money--a Darker Side of Economy. *Indian Journal of Applied Research, 6*.

14. Jain, M. (2017). Making towards a Cashless Economy: Challenges and Opportunities for India. *Indian Journal of Applied Research, 7*.

15. Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011). Target-dependent twitter sentiment classification. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, (pp. 151-160).

16. Krishnan, D., & Siegel, S. (2017). Effects of demonetisation: Evidence from 28 Slum Neighborhoods in Mumbai.

17. Lahiri, A. K., & others. (2016). *Demonetisation, the Cash Shortage and the Black Money.* Tech. Rep.

18. Mahajan, P., & Singla, A. (2017). Effect of demonetisation on financial inclusion in india.

19. Mehta, S., Patel, K., Mehta, K., & others. (2016). *Demonetisation: Shifting Gears From Physical Cash To Digital Cash.* Tech. Rep.

20. Mukhopadhyay, B. (2016). Understanding cashless payments in India. *Financial Innovation, 2*, 27.

21. Pachare, S. M. (2016). Demonetisation: Unpacking the Digital Wallets. *We'Ken-International Journal of Basic and Applied Sciences, 1*, 180-183.

22. Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *Lrec*, *10.*

23. Rajendran, J., & Maggo, V. (2017). Demonetisation: an exuberate sight or plight of indian economy? *Indian Journal of Applied Research, 7*.

24. Rakesh, V. P. (2017). Demonetisation: Endeavor to Script the New Paradigm of Political Economy of India. *PARIPEX-Indian Journal of Research, 6*.

25. Rani, G. (2016). Effects of demonetisation on retail outlets. *IJAR, 2*, 400-401.

26. Rani, S. (2016). Demonetisation: pm Modi's tool to root out black money, terrorist activities from Indian economy.

27. Sarkar, S. (2012). The parallel economy in India: causes, impacts and government initiatives. *Economic Journal of Development Issues, 11*, 124-134.

28. Sharma, C. (2016). Estimating the Size of Black Economy in India.

29. Singh, P., & Singh, V. (2016). The impact of demonetisation on Indian economy.

30. Olson, D. L., & Delen, D. (2008). *Advanced data mining techniques.* Springer Science & Business Media.