Cluster Computing (2019) 22:S14709–S14720
https://doi.org/10.1007/s10586-018-2384-8

Predictive big data analytic on demonetization data using support vector machine

Nattar Kannan[1] · S. Sivasubramanian[1] · M. Kaliappan[2] · S. Vimal[2] · A. Suresh[3]

Received: 30 January 2018 / Revised: 26 February 2018 / Accepted: 6 March 2018 / Published online: 14 March 2018

Abstract
Predictive analytics is the branch of the advanced analytics which makes the user to predict the future events with current statistics. The patterns found in historical and transactional data can be used to identify risks and opportunities for future. Predictive analytics models capture relationships among many factors to assess risk with a particular set of conditions to assign a score. This paper provides predictive analysis on demonetization data using support vector machine approach (PAD-SVM). The proposed PAD-SVM system involved three stages including preprocessing stage, descriptive analysis stage, and prescriptive analysis. The pre-processing stage involves cleaning the obtained data, performing missing value treatment and splitting the necessary data from the tweets. The descriptive analysis stage involves finding the most influential people regarding this subject and performing analytical functionalities. Semantic analysis also is performed to find the sentiment values of the users and to find the compound polarity of each tweet. Predictive analysis is performed to view the current mindset of people and the society reacts to the issue in the current time. This analysis is performed to find out the overall view point of the society and their view may change in the near-future in regarding to the scheme of demonetization as well.

Keywords Descriptive analysis · Predictive analysis · Support vector machine · Sentiment analysis

1 Introduction

Demonetization is the act where current form of money is stopped in circulation or replacement of older currencies or coins with new currencies. Based on the government regulations the measurable steps are being regulated in accordance with RBI to ensue the TAX evasion and to eradicate the combat inflation, corruption and crime.

Predictive analytics [1] is the branch of the advanced analytics which makes the user to predict the future events with current statistics. The patterns found in historical and transactional data can be used to identify risks and opportunities for future. Predictive analytics models capture relationships among many factors to assess risk with a particular set of conditions to assign a score. By successfully applying predictive analytics, the businesses can effectively interpret big data for their benefit. The statistics make use of text analytics with data mining to develop a predictive intelligence based on the relationship in structured and unstructured data [2]. Predictive analytics allows organizations to become proactive, forward looking,

✉ S. Sivasubramanian
ssivasubramaniandce@gmail.com

Nattar Kannan
Kannannattar@gmail.com

M. Kaliappan
kalsrajan@yahoo.co.in

S. Vimal
vimal28.05.1984@gmail.com

A. Suresh
prisu6esh@yahoo.com

[1] Department of CSE, Dhanalakshmi College of Engineering, Chennai, Tamil Nadu, India

[2] Department of Information Technology, National Engineering College, Kovilpatti, Tamil Nadu, India

[3] Department of CSE, Nehru Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India

anticipating outcomes and behaviors based upon the data and not on a hunch or assumptions [3]. The contributions of the proposed scheme crisped as follows:

The proposed PAD-SVM scheme involved three stages including preprocessing stage, descriptive analysis stage, and prescriptive analysis.

- The pre-processing stage involves cleaning the obtained data, performing missing value treatment and splitting the necessary data from the tweets.
- The descriptive analysis stage involves finding the most influential people regarding this subject and performing analytical functionalities.
- Semantic analysis also performed to find the sentiment values of the users and to find the compound polarity of each tweet. Once the polarity scores are calculated, categorize these tweets as "POSITIVE", "NEGATIVE" and "NEUTRAL".
- Finally, the predictive analysis is performed that creates the time-frame for the tweets with the given data, defining the DEPENDENT and INDEPENDENT variables, transforming the data necessary for processing

Through extensive experimental results, we evaluate the performance of the proposed PAD-SVM scheme. The remainder of the paper contains five sections. Section 2 review related work. Section 3 describes our proposed PAD-SVM scheme. Section 4 presents our experimental results and a relevant performance analysis in terms of execution time and classification error. Finally, Sect. 5 presents our conclusions and discusses future direction.

2 Related works

Wilson et al. [2], proposed integrated predictive analytics and social media framework that perform analysis and prediction. This framework used machine learning algorithms to perform predictive analytics tasks, such as feature selection, parameter optimization and result validation.

Jeffery et al. [4] proposed predictive analytics using data mining technique in which data mining techniques were applied on large data sets for prediction. Jimmy and Kolcz [5] proposed polarity classification approach to make classification and enhance accuracy in data sets. Jiang et al., [6] proposed Towards Large-Scale Twitter Mining technique for Drug-Related Adverse Events. It described an approach to find drug users and potential adverse events by analyzing the content of twitter messages utilizing NLP. Bingwei et al. [7] proposed that scalable sentiment classification for big data analysis using Naive Bayes Classifier, Machine learning technologies are widely used in sentiment classification and evaluate the scalability of NBC in large-scale datasets. Cuesta et al., [8] proposed a framework for massive twitter data extraction and analyze data from Twitter's public streams. The framework included a language-agnostic sentiment analysis module, which provides a set of tools to perform sentiment analysis of the collected tweets. Michal and Romanowski [9] proposed a sentiment analysis of twitter data within big data distributed environment for stock prediction and discussed a possibility of making prediction of stock market based on classification of data coming from twitter micro blogging platforms.

Mohit et al. [10], proposed that multi-class tweet categorization using map reduce paradigm, and apache Hadoop framework, Tao et al. [11], proposed sentiment analysis technology and polarity computation of sentiment words. The consumer gets some balance between the price and certain attributes he may concern most. If there are only dozens of reviews, ordinary browsers can just handle them.

Yingyi et al. [12], proposed a new way of sentiment analysis in product evaluation, When purchasing a product for the first time one usually needs to choose among several products with similar characteristics. The best way to choose the most suitable product is to rely upon the opinions of others. The system to be described here collects opinions about hotels from the web, evaluates them, aggregates these evaluations and offers cumulative, easy-to-understand information. Generated information is intended for the possible prospective customer, but also for the hotel managers providing them with additional guidance in future business development.

Maite et al. [13], proposed that sentiment analysis in twitter using machine learning technique, sentiment analysis deals with identifying and classifying opinions or sentiments expressed in source text. Social media is generating a vast amount of sentiment rich data in the form of tweets, status updates, blog posts etc. Sentiment analysis of this user generated data is very useful in knowing the opinion of the crowd. Knowledge base approach and Machine learning approach are the two strategies used for analyzing sentiments from the text.

Tosi et al. [3], proposed that big data from cellular networks: real mobility scenarios for future smart cities and describes a novel use of big data coming from the cellular network of the Vodafone Italy Telco operator to compute mobility patterns for smart cities. These mobility patterns are able to describe different mobility scenarios of the city, starting from how people move around Point Of Interests of the city in real-time. Katkar et al. [14], proposed that real time sentiment analysis of twitter data using Hadoop. Hadoop cluster architecture able to process huge amount of data faster in real time to analysis the sentiment.

Jose et al. [15], proposed that mining twitter big data to predict 2013 Pakistan election winner and analyze the impact of tweets in predicting the winner of the recent 2013 election held in Pakistan. Identify relevant twitter users, pre-process their tweets, and construct predictive models for three

representative political parties which were significantly tweeted. The predictions for last 4 days before the elections showed that party1 emerged as the election winner, which was actually won by party2. The Rapid Miner tool used to experiment with three different standard predictive models.

Wook et al. [16] proposed that big data and predictive analytics methods for modeling and analysis of semiconductor manufacturing processes, Semiconductor manufacturing process generate huge amount of data and harness the value from this data using predictive analytics methods. Kaliappan et al. [17–19], discussed clustering the networks based on the dynamic genetic algorithms. Visual sentimental analysis [20] and Sentiment analytics [21] play a major role in predicting future in hybrid system and twitter data streaming system respectively. Security mechanism [22] was needed to analysis the twitter data. A various enhanced security mechanisms [23] proposed to analysis the security threats and attacks in various domain such as Ad hoc networks, wireless cognitive networks [24]. These security mechanisms may be suitable for analysis the security threats in twitter data. Sudhakar Ilango et al. [25] proposed Artificial Bee Colony approach to select optimal clusters for big data. It imitates the bee behavior to select the clusters. Data distribution, networked and security discussed in Suresh co authors [26–28].

3 Proposed work

The proposed PAD-SVM scheme study the various data obtained from twitter streams which includes the tweets sent by various users and analyzes the present available data to predict the pattern of the data to find its relative future trend. The proposed PAD-SVM system consists of three modules for finding and performing operation on social media data sets. The main scope of the project is to analyze and fetch the twitter IDs of those users whose statuses have been re-tweeted [14] the most by the user whose tweets are being analyzed. First, the system involves collecting the tweets from the social network using the twitter. Then, this consists of standard platform as Hadoop to solve the challenges of big data through map reduce framework where the complete data is mapped to frequent datasets and reduced to smaller sizable data to ease of handling. Finally, it includes analyze the collected tweets and fetching the twitter IDs of those users whose statuses have been re-tweeted the most by the user whose tweets are being analyzed [5]. This system proposes three modules for finding and performing operation on social media data sets [15]. The main scope of the project is to analyze and fetch the twitter IDs of those users whose statuses have been re-tweeted the most by the user whose tweets are being analyzed [6].

3.1 Pre-processing

In the preprocessing stage, the dataset is loaded into the Hadoop file-system. In order to access the files available in the Hadoop, need an interface to connect the HDFS with python application. Pydoop is a python interface to Hadoop that allows the user write applications in pure python [16]. Once the files are accessed in Hadoop using Pydoop, it loads the data in a faster and efficient way. The pandas package used to read the dataset in n-dimensional structure [11]. Table 1 shows the sample data from the dataset

Table 1 Sample twitter data

Id	Text	Favorited	Created	Status source	Retweet count	isRetweet	Retweeted
1	RT @rssurjewala: critical question: was PayTM…	False	23–11–2016 18:40	<a href = "http://twitter.com/download/android"…	331	True	False
2	RT @Hemant_80: did you vote on #Demonetization…	False	23–11–2016 18:40	<a href = "http://twitter.com/download/android"…	66	True	False

Table 2 Processed data with the tweets and users separated from one another

Id	Text	Favorited	Favorite count	Created	Retweet count	isRetweet	Retweeted	Text_new	Users
1	RT @rssurjewala: critical question: Was PayTM…	False	0	23–11–2016 18:40	331	True	False	Critical question: was PayTM informed about #…	RT @rssurjewala
2	RT @Hemant_80: did you vote on #Demonetization…	False	0	23–11–2016 18:40	66	True	False	Did you vote on #Demonetization on Modi surve…	RT @Hemant_80

Algorithm for Data Preprocessing

Input: dataset from Hadoop file System

Output: Processed Data

```
Initialization:
    load twitter-demonetization-data set
    while EOF do
        replace the NaN as zero
    end
    set user  and tweets
Processing:
    do until reach all tweets
        if  re-tweets occurs  then
            split the user from tweets
            split the tweets from user
        else
            split the user as 'other'
            store the existing tweets
    end
```

Load the dataset, to perform the missing value treatment for checking whether the dataset contains any missing values or not. The missing values are filled with zero to avoid errors in processing. Now, split the tweets into re-tweeted users and only the tweets. Split the users by separating them using the semicolon (:) and check whether the tweet contains "RT @" (which denotes the tweet is re-tweeted) [1]. If a tweet is a re-tweeted, the name is entered into dataset. If not, "other" is entered. Split the tweets by separating them using the regex ('(? <=:)(.*)') and enter the first group in them. Table 2 shows the processed data with the tweets and users separated from one another.

Table 3 User list based on Retweet counts and percentage

Users	Retweet count	Retweet percentage
RT @gauravcsawant	541	6.7625
RT @ModiBharosa	539	6.7375
RT @DrKumarVishwas	350	4.3750
RT @rssurjewala	280	3.5000
RT @centerofright	236	2.9500
RT @ashu3page	158	1.9750
RT @kanimozhi	151	1.8875
RT @ShashiTharoor	142	1.7750
RT @Atheist_Krishna	133	1.6625
RT @Joydas	113	1.4125
RT @ippatel	110	1.3750
RT @Joydeep_911	102	1.2750
RT @PIB_India	97	1.2125
RT @DrGPradhan	83	1.0375

3.2 Descriptive analysis

Algorithm for Descriptive Analysis

```
Input: Preprocessed Data
Output: Find Sentiment Type for each tweet
    Initialization:
        Load sentimentIntensityAnalyzer from nltk
        Load WordNetLemmatizer from nltk
        Load tokenize from nltk
        Wid <- WordNetLemmatizer()
        sid<- SentimentIntensityAnalyzer()
        settext_lem and sentiment_compound_polarity
        set sentiment_pos,sentiment_negative and sentiment_neutral
    Processing:
Calculate the Users Based on Number of Retweet
Calculate the Users Based on Their Percentile of Retweet
Initialize text_lem
do until reach all processed tweets
    Strip the tweets
    Get the characters by cleansing the special characters
    Lemmatize the tweet
end
do until reach all processed tweets
Find polarity scores of tweet using Sentiment Intensity Analyzer
Set compound polarity_scores (positive, negative, neutral) to
sentiment_compound_polarity
Set pospolarity_scores of tweets to sentiment_pos
Set negpolarity_scores of tweets to sentiment_negative
Set neupolarity_scores of tweets to sentiment_neutral
end
// Find The Sentiment Type (POSITIVE,NEUTRAL,NEGATIVE)
Set sentiment_type as object
do until reach all sentiment_compound_polarity values
    if sentiment_compound_polarity > 0 then
        set sentiment_type as "POSITIVE"
    else if sentiment_compound_polarity == 0 then
        set sentiment_type as "NEUTRAL"
    else if sentiment_compound_polarity< 0 then
        set sentiment_type as "NEGATIVE"
end
```

The output of pre-processing stage is fed into input of descriptive analysis phase. In order to find the most

influential people regarding the demonetization, ordering the users based on the number of times whose tweets has been re-tweeted [8]. Table 3 shows the top 14 people with most re-tweet count. Also, individual influence on the majority of people could be found by ordering the users based on their re-tweet percentage. Table 3 shows the top 14 people with their re-tweet percentage.

The processed tweets are taken and perform collaborative functions on them. First, remove the unnecessary data in the processed tweets. Second, cleanse the special characters in them. Then, lemmatize the obtained pre-processed data to group the various forms [21] and join all the processed words as a single tweet. This process is repeated for all the tweets.

Sentiment analysis is performed for finding the polarity scores for each tweet using sentiment Intensity Analyzer [9]. The polarity scores for compound, neutral, negative and positive scores has been calculated. Compound polarity classifies the tweets as sentiment values. If the compound polarity is greater than zero, the tweet is represented as "POSITIVE". If it is lesser than zero, the tweet is represented as "NEGATIVE" [8]. If the tweet is equal to zero, the tweet is represented as "NEUTRAL". All these values are categorized as sentiment type. Table 4 shows the processed data along with sentiment type and polarity scores for each tweet.

Table 5 shows the count of each sentiment type (POSITIVE, NEGATIVE and NEUTRAL) while Table 6 shows the percentage of each sentiment type.

3.3 Predictive analysis

Algorithm for Predictive Analysis

Input: Descriptive Data

Output: Predict the Sentiment Type

Initialization:

Load LabelEncoder from sklearn

Load accuracy_score from sklearn.metrics

set minute , hour , date as object

dep_var = 'sentiment_type'

indep_var =

['sentiment_pos','sentiment_negative','sentiment_neutral','Minute','Hour','Date']

Classification:

Minute = get minute value from 'created' in tweets

Hour = get hour value from 'created' in tweets

Date = get date value from 'created' in tweets

list the object type

do until reach all the objtype

transform the objtype to numeric for classification

end

set model as SVC(gamma = 0.001 , C = 100)

fit the model for dep_var with indep_var

predict the model using SVC

find the accuracy score of the model

match the predict_model to the dataset

map the numeric into sentiment_type

do until reach all the sentiment_type

map 0 as 'Positive'

map 1 as 'Neutral'

map 2 as 'Negative

end

The resultant output from the descriptive analysis is taken as input for this stage. First, to declare DEPENDENT and INDEPENDENT variables [10]. The DEPENDENT variables is the target output to perform the analysis on. The INDEPENDENT variables are the columns that are used to support the analysis done on DEPENDENT variables. Also, separate the minute, hour and date from the tweets

Table 4 Descriptive data

Text	Created	Text_lem	Sentiment_compound_polarity	Sentiment_neutral	Sentiment_negative	Sentiment_positive	Sentiment_type
RT @rssurjewala: critical question: Was PayTM…	23–11–2016 18:40	Critical question was PayTM informed about D…	0.1027	0.762	0.110	0.129	POSITIVE
RT @Hemant_80: did you vote on #Demonetization…	23-11-2016 18:40	Did you vote on demonetization on Modi survey…	0.0000	1.000	0.000	0.000	NEUTRAL
RT @gauravcsawant: Rs 40 lakh looted from a ba…	23-11-2016 18:38	Rs lakh looted from a bank in Kishtwar in J…	−0.6249	0.806	0.194	0.000	NEGATIVE

Table 5 Sentiment type in count and percentage

Sentiment type	Count	Percentage
POSITIVE	3068	38.35
NEUTRAL	2568	32.1
NEGATIVE	2364	29.55

Table 6 Sentiment type by time

Sentiment_type	Hour	Count
NEGATIVE	0	23
	1	27
	…	…
	22	12
	23	11
NEUTRAL	0	21
	1	42
	…	…
	22	14
	23	9
POSITIVE	0	32
	1	24
	…	…
	22	14
	23	5

time frame [20].The cryptographic techniques will also be used with the data set to attain the future datas [26].

The transformation of object columns in the dataset into integer types is done using LabelEncoder. LabelEncoder is a utility to help normalize labels such that they can transform non-numerical values into numerical values. Now, the SVC has been set to model. The objective of using SVC is to fit the data, also it is returning the "best fit" hyperplane that categorizes the DEPENDENT and INDEPENDENT data. After getting the hyperplane, can feed the INDEPENDENT data to the classifier to see the predicted output. The accuracy of the model can calculated by using accuracy score to calculate the precision of the model [4].The required predicted set has been obtained which are used to override with the present data. Now, the outcome is in integer type. In order to produce the output in terms of classification, assign the numerical values to their respective types such as 0 to 'POSITIVE', 1 to 'NEUTRAL' and 2 to 'NEGATIVE' [12]. The Table 6 shows the predicted Sentiment type (NEGATIVE, NEUTRAL and POSITIVE) for each hour.

Table 7 shows the predicted sentiment type (POSITIVE, NEGATIVE and NEUTRAL) and the percentage of each predicted sentiment type.

Table 7 Predicted sentiment type in count

Sentiment type	Count	Percentage
NEGATIVE	3207	40.08
NEUTRAL	2545	31.81
POSITIVE	2248	28.11

4 Results and discussions

The proposed PAD-SVM scheme is implemented in Pydoop architecture multimode cluster environment that yield the results. The data analytic process is performed on Demonetization data. This dataset is downloaded from UCI Machine Learning Dataset Repository.

4.1 Descriptive analysis

Figure 1 shows the total count of each sentiment type in bar chart representation. The graph contains the values that are obtained from the descriptive analysis. This shows that the majority of people are in support of the demonetization of 500 and 1000 rupee note with a total count 3068 out of 8000 people. Even though, there are large numbers of people showing their support to this scheme, 2364 out of 8000 have opposing views regarding this issue. The remaining 2568 out of 8000 people have conflicting views about this issue which shows that they are neither negative nor positive but neutral.

Figure 2 shows the percentage of each sentiment type in pie chart representation. The graph also contains the values that are obtained from the descriptive analysis. While the bar chart represents the data in terms of count, we need an overall view regarding the analysis. This graph shows that 38.35% of people are in support of the demonetization of 500 and 1000 rupee note. Even though, there are majority of people are showing their support to this scheme, 29.55% of people have opposing views regarding this issue. The remaining 32.1% of people have conflicting views about this issue which shows that they are neither negative nor positive but neutral.

Figure 3 provides the linear time representation for each sentiment type per hour. The graph represents the sentiment values such as POSITIVE, NEGATIVE and NEUTRAL along with the timeframe. Since our dataset is constrained by a day, the time is embodied in terms of hour. The overall highest peak in the graph is positive which is achieved in noon at a rate of 300 tweets. The highest peak for negative is achieved in the morning while the highest peak for neutral is in the evening.

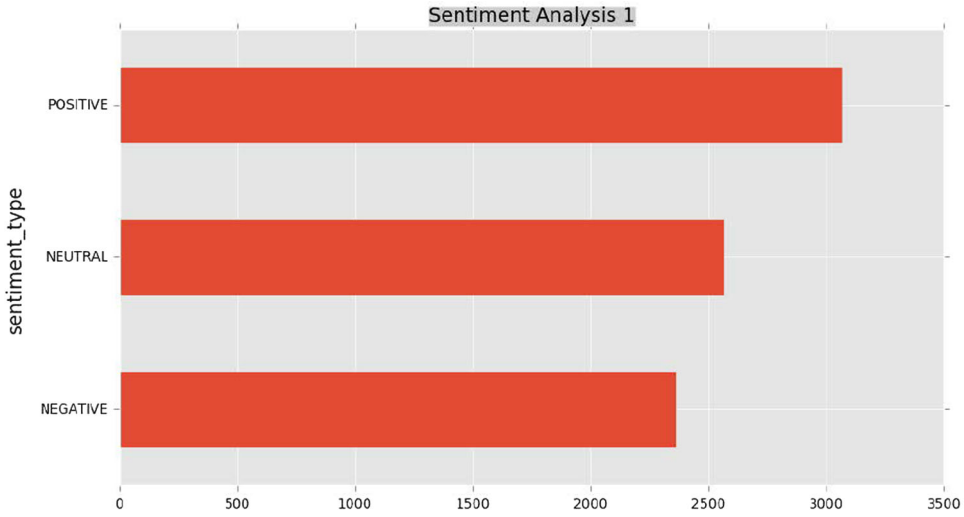

Fig. 1 Sentiment type versus people

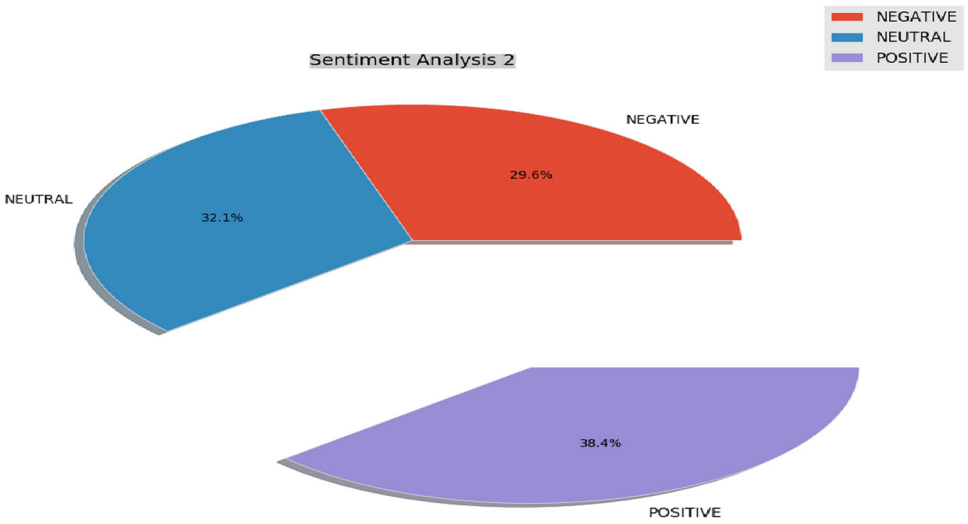

Fig. 2 Sentiment type versus percentage

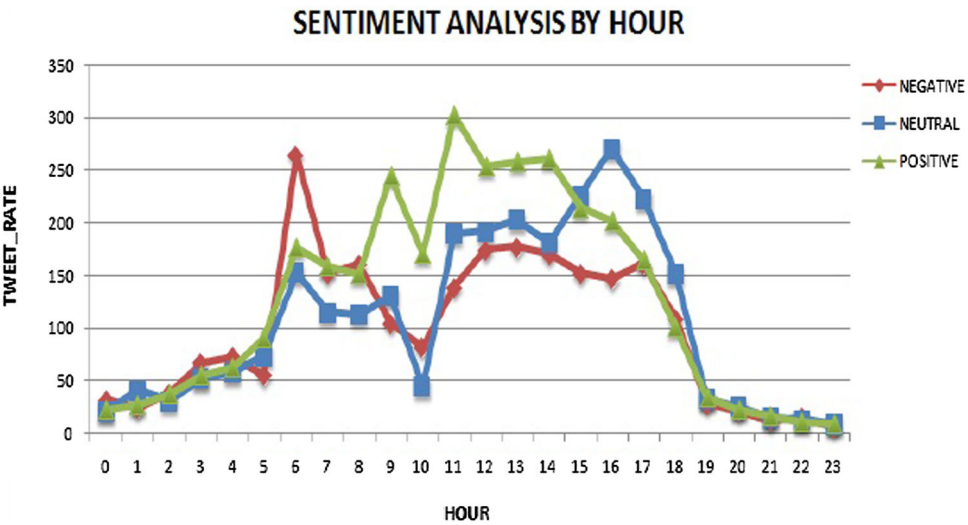

Fig. 3 Sentiment analysis

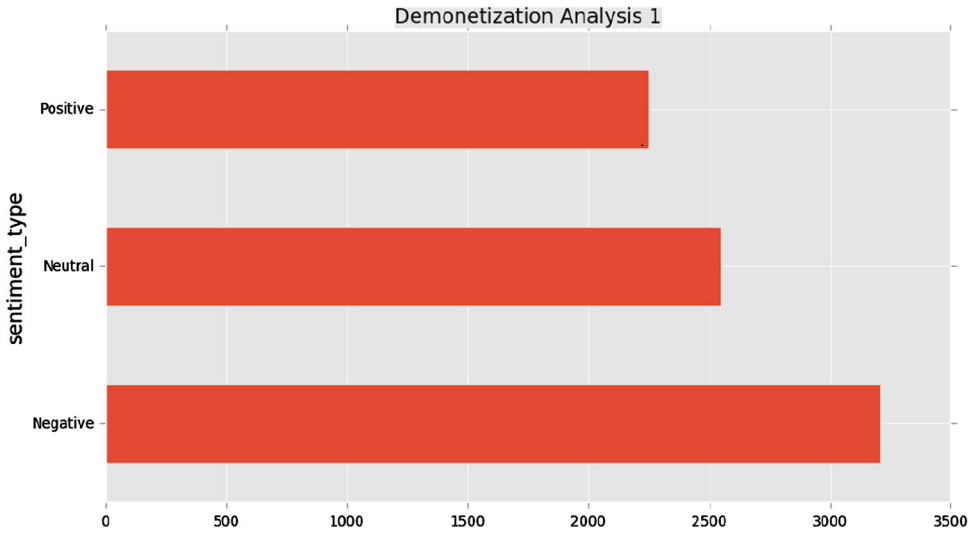

Fig. 4 Predicted sentiment type versus people

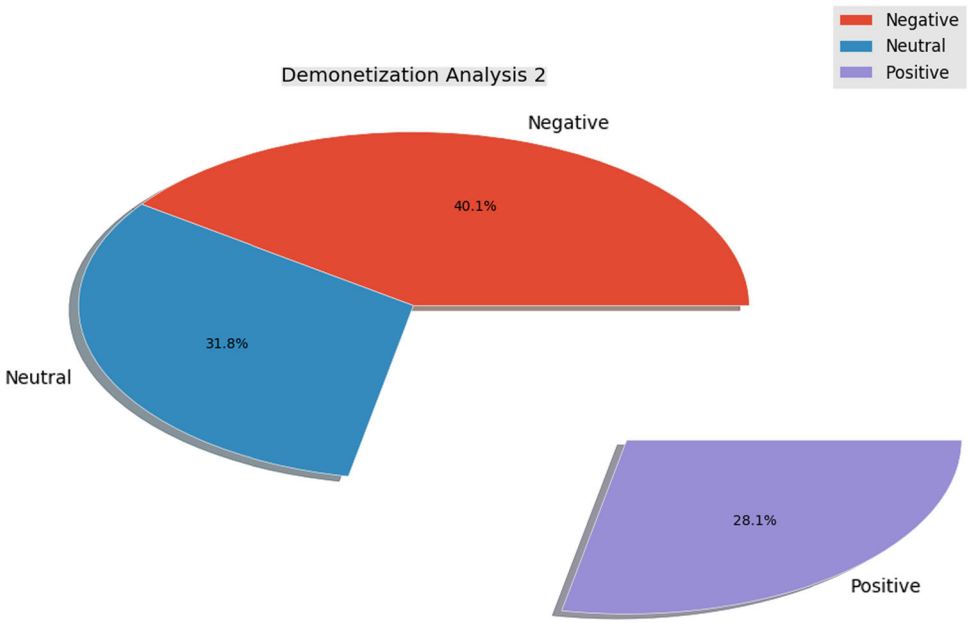

Fig. 5 Predicted sentiment type versus percentage

4.2 Predictive analysis

Figure 4 shows the total count for predicted values of each sentiment type. The graph contains the values that are obtained from the descriptive analysis. This shows that the majority of people who were in support of the demonetization of 500 and 1000 rupee note could decline 3068–2248. There is also rise of opposing views with a count of 2364–3207. These people could have previously supported the demonetization scheme [13]. There is a little change in the NEUTRAL values with a decline from 2568 to 2545 people.

Figure 5 shows the percentage for predicted values of each sentiment type. The graph contains the values that are obtained from the previous analysis. While the Fig. 4 represents the data in terms of count, we need an overall view regarding the analysis. This graph shows that the POSITIVE views regarding the demonetization scheme has been changed with a drop from 38.35 to 28.11%. The people adverse views towards the scheme have grown from 29.55 to 40.08% which is higher than the POSITIVE views that were calculated during the descriptive analysis. There is a very little change in the values of NEUTRAL which is reduced from 32.1 to 31.81%.

Figure 6 provides the linear time representation for predicted values of each sentiment type per hour in count. The graph represents the predicted sentiment values such as POSITIVE, NEGATIVE and NEUTRAL along with the timeframe. Since our dataset is constrained by a day, the predicted values are also limited by this time which is

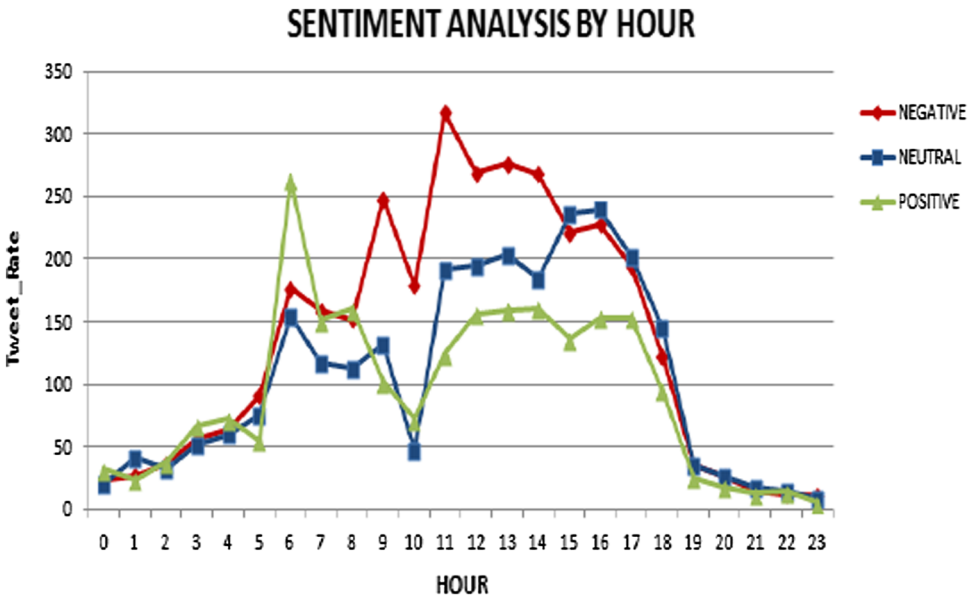

Fig. 6 Predicted sentiment analysis

embodied in terms of hour. The overall highest peak in the graph is NEGATIVE which is achieved around noon at a rate of higher than the POSITIVE values in the descriptive analysis [3]. The highest peak for POSITIVE is achieved in the morning while the highest peak for neutral is in the evening.

5 Conclusion

The effect of demonetization of 500 and 1000 rupees that accounted for 86% of the country's circulating cash has led to very hectic and chaotic events that changed the views of many people in the country. The analysis of the demonetization using twitter data shows descriptive analysis of the current people's view, their sentiment values towards the issue, the people feel about it and how their views might change in the near future. The preprocessing stage involves cleaning the obtained data, performing missing value treatment and splitting the necessary data from the tweets

The descriptive analysis involves finding the most influential people regarding this subject, how much they influence the people and performing analytical functionalities. These analytical functionalities include striping the already processed tweets, cleansing them from special characters, lemmatizing the tweets and using this to find the compound polarity of each tweet. Once the polarity scores are calculated, and categorize these tweets as "POSITIVE", "NEGATIVE" and "NEUTRAL". The graphical representation of the values shows that the 38.35% of people support the idea of demonetization, 32.1% are feeling conflicted about the idea and 29.55% of people oppose the idea of demonetization. The predictive analysis creates the time-frame for the tweets with the given data, defining the DEPENDENT and INDEPENDENT variables, transforming the data for processing. The transformed data is used to fit the SVC model with the DEPENDENT and INDEPENDENT to make it ready for processing. The fitted model is predicted using the SVM technique with INDEPENDENT variables. The accuracy score of the Predicted Model is 96.66%. The predicted model is used to override the existing data to find the predicted data. The graphical representation of the values shows that the 40.08% of people oppose the idea of demonetization, 31.81% are feeling conflicted about the idea and 28.11% of people support the idea of demonetization. From the obtained information, it can be seen that during the initial stage, the majority of people support demonetization. But as the time progresses, the positive views are plummeting and there is an increase in the negative tide which shows that many people who first supported the idea are changing their views.

References

1. Márquez-Vera, C., Morales, C.R., Soto, S.V.: Predicting school failure and dropout by using data mining techniques. IEEE J. Latin-Am. Learn. Technol. **8**(1), 7–14 (2013)
2. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of HLT and EMNLP, vol. 5, pp. 347–354. ACL, New York (2005)
3. Tushar, R., Srivastava, S.: Analyzing stock market movements using twitter sentiment analysis. In: Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), vol. 6. IEEE Computer Society (2012)

4. Jeffrey, D., Ghemawat, S.: MapReduce: simplified data processing on large clusters. Commun. ACM **4**, 107–113 (2008)
5. Jimmy, L., Kolcz, A.: Large-scale machine learning at twitter. In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, vol. 8, pp. 793–804. ACM, New York (2012)
6. Jiang, B., Topaloglu, U., Yu, F.: Towards large-scale twitter mining for drug-related adverse events. In: Proceedings of the 2012 International Workshop on Smart Health and Wellbeing, vol. 2, pp. 25–32. ACM, New York (2012)
7. Bingwei, L., Blasch, E., Chen, Y., Shen, D., Chen, G.: Scalable sentiment classification for big data analysis using Naive Bayes classifier. In: 2013 IEEE International Conference on Big Data, vol. 5, pp. 99–104. IEEE (2013)
8. Cuesta, A., Barrero, D.F.: MD R-Moreno: A framework for massive twitter data extraction and analysis. Malays. J. Comput. Sci. **3**, 50–67 (2014)
9. Michal, S., Romanowski, A.: Sentiment analysis of twitter data within big data distributed environment for stock prediction. In: 2015 Federated Conference on Computer Science and Information Systems, vol. 2, pp. 1349–1354. IEEE (2015)
10. Mohit, T., Gohokar, I., Sable, J., Paratwar, D., Wajgi, R.: Multiclass tweet categorization using map reduce paradigm. Int. J. Comput. Trends Technol. **3**, 78–81 (2014)
11. Tao, C.C., Kim, S.K., Lin, Y.A., Yu, Y.Y., Bradski, G., Ng, A.Y., Olukotun, K.: Map-reduce for machine learning on multicore. NIPS **6**, 281–288 (2006)
12. Yingyi, B.: HaLoop: efficient iterative data processing on large clusters. In: Proceedings of the VLDB Endowment 3.1-2, vol. 6, pp. 285–296 (2010)
13. Maite, T.: Lexicon-based methods for sentiment analysis. Comput. Linguist. **7**, 267–307 (2011)
14. Katkar, V.D., Kulkarni, S.V.: A novel parallel implementation of Naive Bayesian classifier for big data. In: International Conference on Green Computing, Communication and Conservation of Energy, vol. 7, pp. 847–852, ISBN 978-1-4673-6126-2/2013. IEEE (2015)
15. Jose, A.K., Bhatia, N., Krishna, S.: Twitter Sentiment Analysis, vol. 5. National Institute of Technology Calicut, IEEE, Calicut (2010)
16. Wook, M., Hani Yahaya, Y., Wahab, N.: Predicting NDUM student's academic performance using data mining techniques. In: Second International Conference on Computer and Electrical Engineering, vol. 3, pp. 357–361. IEEE (2009)
17. Kaliappan, M., Augustine, S., Paramasivan, B.: Enhancing energy efficiency and load balancing in mobile adhoc network using dynamic genetic algorithms. J. Netw. Comput. Appl. **73**, 35–43 (2016)
18. Kaliappan, M., Mariappan, E., Prakash, M.V., Paramasivan, B.: Load balanced clustering technique in MANET using genetic algorithms. Defence Sci. J. **66**(3), 251–258 (2016). https://doi.org/10.14429/dsj.66.9205
19. Subbulakshmi, P., Vimal, S.: Secure data packet transmission in MANET using enhanced identity-based cryptography. Int. J. New Technol. Sci. Eng. **3**(12), 35–42 (2016)
20. Ghiassi, M., Skinner, J., Zimbra, D.: Twitter brand sentiment analysis: a hybrid system using n-gram analysis and dynamic artificial neural network. Expert Syst. Appl. **40**(16), 6266–6282 (2013)
21. Hao, M., Rohrdantz, C., Janetzko, H., Dayal, U., Kiem, D.A., Haug, L.E., Hsu, M.C.: Visual sentiment analysis on twitter data streams. In: IEEE Symposium on Visual Analytics Science and Technology, vol. 20, October 23, Providence, RI, USA (2014)
22. Kaliappan, M., Paramasivam, B.: Enhancing secure routing in mobile ad hoc networks using a dynamic Bayesian signalling game model. Comput. Electr. Eng. **41**, 301–313 (2015)
23. Mariappan, E., Kaliappan, M., Vimal, S.: Energy efficient routing protocol using Grover's searching algorithm for MANET. Asian J. Inf. Technol. **15**, 4986–4994 (2016). https://doi.org/10.3923/ajit.2016.4986.4994
24. Vimal, S., Kalaivani, L., Kaliappan, M.: Collaborative approach on mitigating spectrum sensing data hijack attack and dynamic spectrum allocation based on CASG modeling in wireless cognitive radio networks. Clust. Comput. (2017). https://doi.org/10.1007/s10586-017-1092-0
25. Sudhakar Ilango, S., Vimal, S., Kaliappan, M., Subbulakshmi, P.: Optimization using Artificial Bee Colony based clustering approach for big data. Clust. Comput. (2018). https://doi.org/10.1007/s10586-017-1571-3
26. Suresh, A., Varatharajan, R.: Competent resource provisioning and distribution techniques for cloud computing environment. Cluster Comput. (2017). https://doi.org/10.1007/s10586-017-1293-6
27. Chinnasamy, A., Sivakumar, B., Selvakumari, P., Suresh, A.: Minimum connected dominating set based RSU allocation for smartCloud vehicles in VANET. Cluster Comput. (2018). https://doi.org/10.1007/s10586-018-1760-8
28. Suresh, A., Reyana, A., Varatharajan, R.: CEMulti-core architecture for optimization of energy over heterogeneous environment with high performance smart sensor devices. Wirel. Pers. Commun. (2018). https://doi.org/10.1007/s11277-018-5504-0

Nattar Kannan born in Kayalpatnam, Tamil Nadu, India on September 17, 1979. He completed B.E. degree in Computer Science and Engineering under Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India in 2002. He obtained M.E. and Ph.D. degree in Computer Science and Engineering under Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India. He works as in Department of Computer Science and Engineering, Dhanalakshmi College of Engineering, Chennai, Tamil Nadu, India. His research interests include Wireless Ad-Hoc networks.

S. Sivasubramanian received his Ph.D. degree in Information and Communication Engineering from Anna University, Chennai, Tamil Nadu, India. He works as an Professor and Head in Department of Computer Science and Engineering, Dhanalakshmi College of Engineering, Chennai, Tamil Nadu, India. He has wide publications in SCI, Scopus indexed Journals & Conferences. He is a member in various professional bodies. Currently, his research interests include, Wireless Ad-Hoc networks and cluster computing.

M. Kaliappan received his Ph.D. degree in Information and Communication Engineering from Anna University, Chennai, Tamil Nadu, India. He works as an Associate Professor in Department of Information Technology, National Engineering College, Kovilpatti, Tamil Nadu, India. He has wide publications in SCI, Scopus indexed Journals & Conferences. He is a member in various professional bodies and organized varied funded programs. Currently, his research interests include Game Modeling, Big Data, Analytic computing and Wireless Ad-Hoc networks.

S. Vimal is currently working as an Assistant Professor (Senior Grade) in Department of Information Technology, National Engineering College, Kovilpatti, Tamil Nadu. He has around 12 years of teaching experience, EMC certified Data science Associate and CCNA certified professional too. He is a member in various professional bodies and organized varied funded programs. He has wide publications in the highly impact journals in the area of networking and security issues. His areas of interest include Game Modeling, Cognitive radio networks, network security and Big data Analytics.

A. Suresh works as the Professor & Head, Department of the Computer Science and Engineering in Nehru Institute of Engineering & Technology, Coimbatore, Tamil Nadu, India. He has been nearly two decades of experience in teaching and his areas of specializations are Data Mining, Artificial Intelligence, Image Processing, Multimedia and System Software. He has published 25 papers in International journals. He has published more than 40 papers in National and International Conferences. He has served as a reviewer for Springer, Elsevier, and Inderscience journals. He is a member of ISTE, IACSIT, IAENG, MCSTA, MCSI, and Global Member of Internet Society (ISOC). He has organized several National Workshop, Conferences and Technical Events. He is regularly invited to deliver lectures in various programmes for imparting skills in research methodology to students and research scholars. He has published three books, in the name of Data structures & Algorithms, Computer Programming and Problem Solving and Python Programming in DD Publications, Excel Publications and Sri Maruthi Publisher, Chennai, respectively.