PREDICTIVE ANALYSIS ON DEMONETIZATION DATA USING SUPPORT VECTOR MACHINE TECHNIQUE

KALIAPPAN M[1], MARIAPPAN E[2], RAMNATH M[3], KARPAGAVALLI C[4], ANGEL HEPZIBAH R[5] , VIMAL S[6]

[1]Professor, Department of Artificial Intelligence and Data Science, Ramco Institute of Technology, Rajapalayam, Tamil Nadu, India
[2]Associate Professor, Department of Artificial Intelligence and Data Science, Ramco Institute of Technology, Rajapalayam, Tamil Nadu, India
[3,4,5]Asssitant Professor, Department of Artificial Intelligence and Data Science, Ramco Institute of Technology, Rajapalayam, Tamil Nadu, India
[6] Professor, Department of Artificial Intelligence and Data Science, Sri Eshwar College of Engineering, Coimbatore, Tamil Nadu, India
E-mail: [1]kalsrajan@yahoo.co.in, [2] mapcse.e@gmail.com, [3]ramnath25@gmail.com, [4]ckvalli@gmail.com, [5]rangelhepzibah@gmail.com and [6]svimalphd@gmail.com

ABSTRACT

The Unexpected announcement of demonetization of 500 and 1000-rupee note caused chaos in the cash-dependent economy which resulted in a lot of small-scale industries and entrepreneurs who lost their livelihood due to these events. However, the main intent of this Scheme is to eradicate black-market money, to combat inflation and move towards the betterment of our society. There have been a lot of mixed views towards this issue. In this paper, we develop predictive Analysis on Demonetization Data using SVM approach (PAD-SVM). The proposed system involved three stages including preprocessing stage, descriptive analysis stage, and descriptive analysis. The Preprocessing 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. These analytical functionalities include striping the already processed tweets, cleansing them from special characters, lemmatizing the tweets. Semantic analysis is 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”. Descriptive analysis is performed to view the current mindset of people and the society reacts to the issue in the current time. Predictive analysis is performed to predict the mindset of people which may change. This analysis is performed to find out the overall viewpoint 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 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 with new notes or coins. Sometimes, a country completely replaces the old currency with new currency. There are multiple reasons why nations demonetize their local units of currency such as combat inflation, combat corruption and crime (counterfeiting, tax evasion), discourage a cash-dependent economy.

In 2016, the Indian government decided to demonetize the 500- and 1000- rupee notes, the two biggest denominations in its currency system; these notes accounted for 86% of the country’s circulating cash. With little warning, India's Prime Minister Narendra Modi announced to the citizenry on Nov 8 that those notes were worthless, effective immediately – and they had until the end of the year to deposit or exchange them for newly introduced 2000 rupee and 500-rupee bills. Chaos ensued in the cash-dependent economy (some 78% of all Indian customer transactions are in cash), as long, snaking lines formed outside ATMs and banks, which had to shut down for a day. The new rupee notes have different specifications, including size and thickness, requiring re-calibration of ATMs: only 60% of the country's 200,000 ATMs were operational. Even those dispensing bills of lower denominations faced shortages. The government's restriction on daily withdrawal amounts added to the misery, though a waiver on transaction fees did help a bit. The government's goal (and rationale for the abrupt announcement) was to combat India's thriving underground economy on several fronts: eradicate counterfeit currency, fight tax evasion (only 1% of the population pays taxes), eliminate black money gotten from money laundering and terrorist-financing activities, and to promote a cashless economy. Individuals and entities with huge sums of black money gotten from parallel cash systems were forced to take their large-denomination notes to a bank, which was by law required to acquire tax information on them. If the owner could not provide proof of making any tax payments on the cash, a penalty of 200% of the owed amount was imposed.

Predictive analytics is the branch of the advanced analytics which is used to make predictions about unknown future events. Predictive analytics uses many techniques from data mining, statistics, modeling, machine learning, and artificial intelligence to analyze current data to make predictions about future. It uses a number of data mining, and analytical techniques to bring together the management, information technology, and modeling business process to make predictions about future. 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, or weightage. By successfully applying predictive analytics the businesses can effectively interpret big data for their benefit. The data mining and text analytics along with statistics, allows the business users to create predictive intelligence by uncovering patterns and relationships in both the structured and unstructured data. The data which can be used readily for analysis are structured data, examples like age, gender, marital status, income, sales etc. Unstructured data are textual data in call center notes, social media content, or other type of open text which need to be extracted from the text, along with the sentiment, and then used in the model building process. Predictive analytics allows organizations to become proactive, forward looking, anticipating outcomes and behaviors based upon the data and not on a hunch or assumptions. Prescriptive analytics goes further and suggest actions to benefit from the prediction and also provide decision options to benefit from the predictions and its implications.

2. RELATED WORKS

Steffen Koch, Thomas Ertl, and Ross Maciejewski [1], proposed that integrating predictive analytics and social media, in this paper, a framework for social media integration, analysis and prediction is presented. This framework consists of tools for extracting, analyzing and modeling trends across various social media platforms. This system integrates unstructured data from twitter and you tube with curated data from the Internet Movie Database. These software packages and tools provide a variety of machine learning algorithms that can be used for predictive analytics tasks, such as feature selection, parameter optimization and result validation. This paper presents an interactive framework integrating social media and predictive analytics, and the presentation of a talk aloud study that discusses design successes, pitfalls and potential future directions. It allows for the quick integration of structured and unstructured data sources, focusing on box office predictions as our example domain. The results were validated through case studies and user studies, which have demonstrated that such a tool can quickly enable non-domain experts to be competitive with domain experts in a given area.

Hina Gulati [2] proposed that predictive analytics using data mining technique, prediction can be done by using data mining techniques on large data sets. Data mining is a broad concept that consists of series of steps. Firstly, data is pre-processed and then mining techniques are applied. Results from mining techniques are evaluated and interpreted. Main objective of this seminar is to use prediction method for educational data mining. Reason for choosing education domain for predictive analytics is the availability of data predicting student's dropout reasons can be difficult task due to multiple factors that can affect the decision. In preprocessing step feature selection algorithms are used to identify features that will affect the prediction process the most. After preparing data for mining classification algorithms are applied and by analysis of decision tree and induction rules we get the prediction model that is tested on test data can help to find useful knowledge. Result obtained from such models can help teachers and management to identify the problem areas and reasons that affect dropout the most. We have considered three cases and accuracy when compared for classification in all three cases will lead to understanding that which is most effective way to analyze student performance and help in identifying reasons for drop-out.

Lin, Jimmy, and Alek Kolcz [3] proposed that Large-Scale Machine Learning at Twitter, this paper presents a case study of twitter's integration of machine learning tools into its existing Hadoop-based, Pig-centric analytics platform. The core of this work lies in recent Pig extensions to provide predictive analytics capabilities that incorporate machine learning, focused specifically on supervised classification. It provides a base-line for classification accuracy from content, given only large amounts of data. The data set involves a test set consisting of one million English tweets with emoticons from Sept. 1, 2015, at least 20 characters in length. The test set was selected to contain an equal number of positive and negative examples. In preparing both the training and test sets, emoticons are removed. Their machine learning framework consists of two components: a core Java library and a layer of lightweight wrappers that expose functionalities in Pig. A Pig script was written for training binary sentiment polarity classifiers. The script processes tweets, separately filtering out those containing positive and negative emotions, which are combined together to generate the final training set. Results of the polarity classification experiments showed accuracy in the range 77% to 82% with varying data set size.

Bian, Jiang, UmitTopaloglu, and Fan Yu[4] proposed that Towards Large-Scale Twitter Mining for Drug-Related Adverse Events. The authors describe an approach to find drug users and potential adverse events by analyzing the content of twitter messages utilizing NLP and to build SVM classifiers. The data set used is a collection of over 2 billion tweets collected from May 2009 to October 2010, from which they try to identify potential adverse events caused by drugs of interest. The collected stream of tweets was organized by a timeline. The raw twitter messages were crawled using the twitter's user timeline API that contains information about the specific tweet and the user. Two-class SVM was used for the purpose of classification. Evaluation of the SVM was done using parameters such as, the Area under the Curve (AUC) value, and the Receiver operating characteristic (ROC) curve. The ROC curve using the mean values of the 1000 iterations was drawn. The prediction accuracy on average over the 1000 iterations was evaluated to 0.74 and the mean AUC value is 0.82.

[5] Liu, Bingwei, Erik Blasch, Yu Chen, Dan Shen, and Genshe Chen. In Big Data, 2013 IEEE International Conference on, pp. 99-104. IEEE, 2013, proposed that scalable sentiment classification for big data analysis using Naive Bayes Classifier, Machine learning technologies are widely used in sentiment classification because of their ability to "learn" from the training dataset to predict or support decision making with relatively high accuracy. In this paper, the authors evaluate the scalability of NBC in large-scale datasets. They have presented a simple and complete system for sentiment mining on large datasets using a Naive Bayes classifier with the Hadoop framework. They have demonstrated that NBC is able to scale up to analyze the sentiment of millions movie reviews with increasing throughput. The raw data comes from large sets of movie reviews collected by research communities. In their experiments, they use two datasets: the Cornell University movie review dataset3 and Stanford SNAP Amazon movie review dataset4. The Cornell dataset has 1000 positive and 1000 negative reviews. The Amazon movie review dataset is organized into eight lines for each review, with additional information such product identification (ID), user ID, profile Name, score, summary etc. They tested their code on Cornell dataset and resulted in an 80.85% average accuracy. To test the scalability of Naive Bayes classifier, the size of dataset in their experiment varies from one thousand to one million reviews in each class.

[6] AlvaroCuesta, David F., and Maria D. R-Moreno. In Malaysian Journal of Computer Science, pp 50-67 (2014), proposed that a framework for massive twitter data extraction and analysis. The authors propose an open framework to automatically collect and analyze data from Twitter's public streams. This is a customizable and extensible framework, so researchers can use it to test new techniques. The framework is complemented with a language-agnostic sentiment analysis module, which provides a set of tools to perform sentiment analysis of the collected tweets. Most tools in the framework are implemented in Python, but the Classifier and Tester web interfaces run on NodeJS and are programmed in Coffee Script (a language which can be pre-processed into JavaScript). The chosen backend database is MongoDB. Classification was done according to three classes, "positive", "negative" and "neutral". The conclusion is that the best trainers had 1-grams included and a minimum score between 2 and 4.

[7] Skuza, Michal, and Andrzej Romanowski. In Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on, pp. 1349-1354. IEEE, 2015, proposed that sentiment analysis of twitter data within big data distributed environment for stock prediction. This paper discusses a possibility of making prediction of stock market basing on classification of data coming from twitter micro blogging platform. Twitter messages are retrieved in real time using twitter Streaming API. Tweets were collected over 3 month's period from 2nd January 2013 to 31st March 2013. It was specified in the query that tweets have to contain name of the company. After pre-processing each message was saved as bag of words model – a standard technique of simplified information representation used in information retrieval. Polarity mining is a part of sentiment in which input is classified either as positive or negative. Prediction of future stock prices is performed in this work by combining results of sentiment classification of tweets and stock prices from a past interval. Taking into consideration large volumes of data to be classified and the fact they are textual, Naïve Bayes method was chosen due to its fast-training process even with large volumes of training data and the fact that is it is incremental. Considered large volumes of data resulted also in decision to apply a map reduce version of Naïve Bayes algorithm.

[8] Tare, Mohit, IndrajitGohokar, Jayant Sable, DevendraParatwar, and RakhiWajgi. In International Journal of Computer Trends and Technology. pp 78 - 81 (2014), proposed that multi-class tweet categorization using map reduce paradigm, the authors have proposed strategy that uses apache Hadoop framework, an open source java framework, which relies on map – reduce paradigm and a HDFS to process data. The proposed Map – Reduce strategy for classification of tweets using NBC relies on two map-reduce passes. We have used the twitter4j library to gather tweets which internally uses twitter REST API. The twitter4j library requires OAuth support to access the API. Twitter uses OAuth to provide authorized access to its API. The final step after preprocessing of tweets is the labeling of tweets based on categories namely politics, sports and technology. Then final reducer calculates the final probability of each category to which the tweet may belong to and outputs the predicted category and its probability value as key-value pair. [9] Min Wang, Hanxiao Shi, IEEE (2016), proposed that research on sentiment analysis technology and polarity computation of sentiment words, in this way, the consumer can get 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. In this experiment, the accuracy of orientation value computation for new arrived sentiment words was mainly evaluated. The range of orientation value is from -1 to + 1. + 1 is the highest commendation, and -1 is the largest derogation. The deviation threshold of orientation value of artificial judgment for 0.3 on the condition of correct polarity judgment is allowed. Due to constant deviation between automatic computation and manual annotation, the deviation threshold of orientation value on the condition of correct polarity judgment is allowed.

[10] Lada Banic, Ana Mihanovic, Marko Brakus, May 20 – 24, 2013, opatija, Croatia, IEEE, proposed thatusing big data and 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. Evaluation system involved evaluation of terms and phrases with the help of grades from 1 to 5, where 1 referred to bad and 5 referred to excellent. Each term or phrase recognized in the review was evaluated according to the specification in the dictionary. Average grade for each single review was obtained. The total grade for each hotel was calculated as average grade of all reviews aggregated on hotel level.

[11] Neethu M S, Rajasree R, IEEE (2015), 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. In this paper, we try to analyze the twitter posts about electronic products like mobiles, laptops etc., and using Machine Learning approach. By doing sentiment analysis in a

specific domain, it is possible to identify the effect of domain information in sentiment classification. A new feature vector for classifying the tweets as positive, negative and extract peoples' opinion about products was presented. SVM classifier uses large margin for classification. It separates the tweets using a hyper plane. SVM uses the discriminative function defined as g(X) = wT _(X) + b (2) where 'X' is the feature vector, 'w' is the weights vector and 'b' is the bias vector, () is the nonlinear mapping from input space to high dimensional feature space. 'w' and 'b' are learned automatically on the training set. Here we used a linear kernel for classification. It maintains a wide gap between two classes.

[12] Davide Tosi, Stefano Marzorati, IEEE (2014), proposed that big data from cellular networks: real mobility scenarios for future smart cities, in this paper, we describe 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. These mobility patterns can be exploited by Policy makers to improve the mobility in city or by Navigation Systems and Journey Planners to provide final users with accurate travel plans. The paper discusses five main new mobility patterns and their experimental validation in real industrial setting and for Milan metropolitan city.

[13] Sunil B. Mane, YashwantSawant, SaifKazi, VaibhavShinde, IEEE (2013), proposed that real time sentiment analysis of twitter data using Hadoop, twitter, one of the largest social media site receives tweets in millions every day. This huge amount of raw data can be used for industrial or business purpose by organizing according to our requirement and processing. This paper provides a way of sentiment analysis using Hadoop which will process the huge amount of data on a Hadoop cluster faster in real time.In our approach we focused more on the speed of performing analysis than its accuracy i.e. performing sentiment analysis on big data which is achieved by splitting the various modules of data in following steps and collaborating with Hadoop for mapping it onto different machines part of speech tagged using opennlp. This tagging is used for following various purposes.

i) Stop words removal: The stop words like a, an, this which are not useful in performing the sentiment analysis are removed in this phase. Stop words are tagged as _DT in Opennlp. All the words having this tag are not considered.

ii) Unstructured to structured: Twitter comments are mostly unstructured i.e. 'aswm' is written 'awesome', 'happyyyyyy' to actually 'happy'. Conversion to structured is done by dynamic data records of unstructured to structure and vowels adding.

iii) Emoticons: These are most expressive method available for opinion. The emoticons symbolic representation is converted in to words at this stage i.e. _ to happy.

[14] Dr. Tariq Mahmood, Tasmiyah Iqbal, Farnaz Amin, WajeetaLohanna, AtikaMustafa, IEEE (2013), proposed that mining twitter big data to predict 2013 Pakistan election winner, In this paper, we 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, i.e., Pakistan Tehreek-e-Insaaf (PTI), Pakistan Muslim League Nawaz (PMLN), and MuttahidaQaumi Movement (MQM). The predictions for last four days before the elections showed that PTI will emerge as the election winner, which was actually won by PMLN. We used the Rapid Miner tool to experiment with three different standard predictive models, i.e., CHAID decision tree, Naïve Bayes and SVM. CHAID is well suited for the analysis of larger datasets. It builds non-binary trees where each non-terminal node identifies a split condition on attributes using the Chi squared test, to yield optimum prediction of the label. The naive Bayes is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. Finally, SVM is a non-probabilistic, linear, and a binary predictor which maps a prediction problem into a high dimensional space to determine the support vectors.

[15] Sujata Butte and SainathPatil, IEEE (2015), proposed that big data and predictive analytics methods for modeling and analysis of semiconductor manufacturing processes, Semiconductor manufacturing fabs generate huge amount of data. The big data approaches of data management have increased speed, quality and accessibility of the data. This paper discusses harnessing value from this data using predictive

analytics methods. Various aspects predictive analytics in the context of semiconductor manufacturing are discussed. The limitations of standard methods of analysis and the need to adopt robust methods of modeling and analysis are highlighted. The robust prediction modeling method is implemented on wafer sensor data resulting in improved prediction ability of wafer quality characteristics.

3. PROPOSED WORK

The proposed scheme studies the various data obtained from twitter streams (which includes the tweets sent by various users) and analyze the present available data to predict the pattern of the data to find its relative future trend.

The proposed system consists of three modules for finding and performing operation on social media data sets. The main scope of the project is to analyzing and fetching the twitter IDs of those users whose statuses have been re-tweeted 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. This system proposes three modules for finding and performing operation on social media data sets. The main scope of the project is to analyzing and fetching the twitter IDs of those users whose statuses have been re-tweeted 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.

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

Pre-Processing

In the preprocessing stage, load the dataset 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. 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. Table 5.1 shows the sample data from the dataset

Algorithm for Data Preprocessing

```
Input: data.csv 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 occur then
   split the user from tweets
   split the tweets from user
else
   split the user as ‘other’
    store the existing tweets
end
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). 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 5.2 shows the processed data with the tweets and users separated from one another.

Id	Text	favorited	FavoriteCount	created	RetweetCount	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

4. 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 preprocessing stage is fed into input of descriptive analysis phase. In order to find the most influential people regarding the demonetization, order the users based on the number of times their tweets has been re-tweeted. Table 5.3a shows the top 14 people with most re-tweet count. Also, find out their influence on the majority of people by ordering the users based on their re-tweet percentage. Table 5.3b shows the top 14 people with their re-tweet percentage.

TABLE 4.3a: USER LIST BASED ON RETWEET COUNTS And PERCENTAGE

Users	Retweet Count	Re-tweet 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

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 preprocessed data using the Word Net Lemmatizer to group the various forms. 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. The polarity scores for compound, neutral, negative and positive scores have been calculated. Classify the tweets as sentiment values by using the compound polarity. If the compound polarity is greater than zero, the tweet is “POSITIVE”. If it is lesser than zero, the tweet is “NEGATIVE”. If the tweet is equal to zero, the tweet is “NEUTRAL”. All these values are categorized as sentiment type. Table 5.4 shows all the processed data along with sentiment type and polarity scores for each tweet.

TABLE 4.4: 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 4.5a: SENTIMENT TYPE IN COUNT AND PERCENTAGE

SENTIMENT TYPE	COUNT	PERCENTAGE
POSITIVE	3068	38.35
NEUTRAL	2568	32.1
NEGATIVE	2364	29.55

Table 5.5a shows the count of each sentiment type (POSITIVE, NEGATIVE and NEUTRAL) while Table 5.5b shows the percentage of each sentiment type.

5. 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. The DEPENDENT variables is the target output to perform the analysis on. The INDEPENDENT variables are the column that are used to support the analysis done on DEPENDENT variables. Also, separate the minute, hour and date from the tweet's timeframe.

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.

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'.

Table 4.6 shows the predicted Sentiment type (NEGATIVE, NEUTRAL and POSITIVE) for each hour

Table 4.6: 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

TABLE 4.7a: Predicted Sentiment Type in Count

SENTIMENT TYPE	COUNT	PERCENTAGE
NEGATIVE	3207	40.08
NEUTRAL	2545	31.81
POSITIVE	2248	28.11

Table 5.7a shows the count of predicted sentiment type (POSITIVE, NEGATIVE and NEUTRAL) while Table 5.7b shows the percentage of each predicted sentiment type.

7. RESULTS AND DICUSSIONS

DECSRIPTIVE ANALYSIS

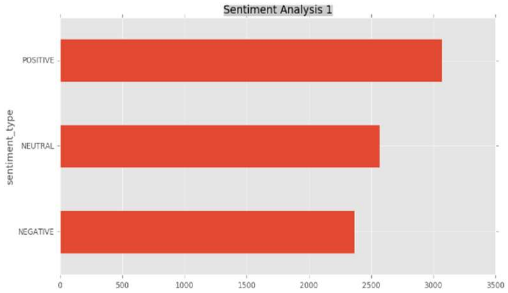

Fig 5.1: Sentiment Type In Bar Chart

Fig 5.1 shows the total count of each sentiment type in bar chart representation. The contents for this graph are taken from Table 5.5a

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.

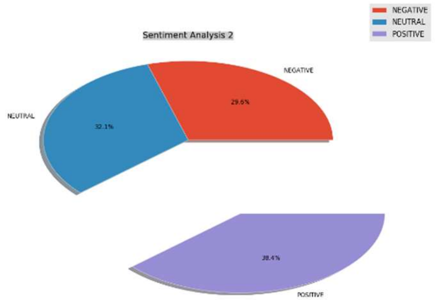

Fig5.2: Sentiment Type in Pie Chart

Fig 5.2 shows the percentage of each sentiment type in pie chart representation. The contents for this graph are taken from Table 5.5b

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.

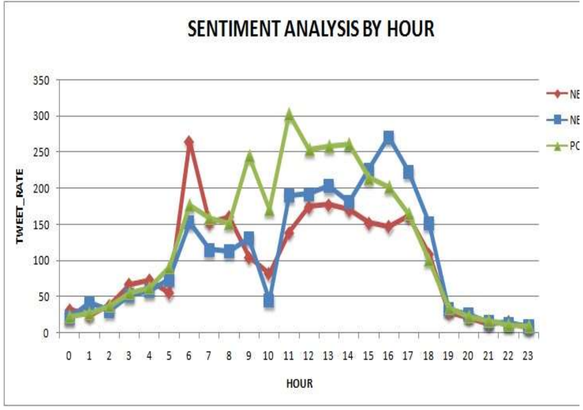

Fig 5.3: Sentiment By Hour in Count

Fig 5.3a provides the linear time representation for each sentiment type per hour in count

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.

8. 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 necessary 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 opposethe 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]. T. Wilson, j. Wiebe and p. Hoffmann, "recognizing contextual polarity in phrase-level sentiment analysis," in proceedings of hlt and emnlp. Acl, vol. 5, (2005), pp. 347–354

[2]. C. C. Tao, s. K. Kim, y. A. Lin, y. Y. Yu, g. Bradski, a. Y. Ng and kunleolukotun, "map-reduce for machine learning on multicore", in nips, vol. 6, (2006), pp. 281-288.

[3]. L. Jimmy, and a. Kolcz, "large-scale machine learning at twitter", in proceedings of the 2012 acm sigmod international conference on

management of data, acm, vol. 8(2012), pp. 793-804.

[4]. B. Jiang, u. Topaloglu and f. Yu, “towards large-scale twitter mining for drug-related adverse events”, in proceedings of the 2012 international workshop on smart health and wellbeing, acm, vol. 2(2012), pp. 25-32.

[5]. L. Bingwei, e. Blasch, y. Chen, d. Shen and g. Chen, “scalable sentiment classification for big data analysis using naive bayes classifier”, in big data, 2013 ieee international conference on, ieee, vol. 5(2013), pp. 99-104.

[6]. A. Cuesta, david f. And maría d. R-moreno, “a framework for massive twitter data extraction and analysis”, in malaysian journal of computer science, vol. 3 (2014), pp. 50-67.

[7]. S. Michal and a. Romanowski, “sentiment analysis of twitter data within big data distributed environment for stock prediction”, in computer science and information systems (fedcsis), 2015 federated conference on, ieee, vol. 2(2015), pp. 1349-1354.

[8]. T. Mohit, i. Gohokar, j. Sable, d. Paratwar and r. Wajgi, “multi-class tweet categorization using map reduce paradigm”, in international journal of computer trends and technology. Vol. 3 (2014), pp. 78-81.

[9]. D. Jeffrey and s. Ghemawat, “mapreduce: simplified data processing on large clusters”, communications of the acm 51.1, vol. 4 (2008), pp. 107-113.

[10]. B. Yingyi, “haloop: efficient iterative data processing on large clusters”, proceedings of the vldb endowment 3.1-2, vol. 6 (2010), pp. 285-296.

[11]. T. Maite, “lexicon-based methods for sentiment analysis”, computational linguistics 37.2, vol. 7(2011), pp. 267-307.

[12]. R. Tushar and s. Srivastava, “analyzing stock market movements using twitter sentiment analysis”, proceedings of the 2012 international conference on advances in social networks analysis and mining (asonam 2012). Ieee computer society, vol. 6 (2012).

[13]. V. D. Katkar, s. V. Kulkarni, “a novel parallel implementation of naive bayesian classifier for big data”, international conference on green computing, communication and conservation of energy, 978-1-4673-6126-2/2013 ieee, vol. 7 (2015) pp. 847-852.

[14]. a. K. Jose, n. Bhatia, and s. Krishna, “twitter sentiment analysis”. National institute of technology calicut, ieee, vol. 5 (2010).

[15]. M. Wook, y. Hani yahaya, n. Wahab, “predictingndum student’s academic performance using data mining techniques”, ieee, second international conference on computer and electrical engineering, vol. 3 (2009),pp.357361

[16]. C. Márquez-vera, c.r.morales, ands.v.soto,”predicting school failure and dropout byusing data mining techniques”, ieee journal oflatin-american learning technologies, vol. 8, no. 1,february 2013, pp.7-14.

[17]. M. Hao, c. Rohrdantz, h. Janetzko, u. Dayal, d. A. Kiem, l.e.haug,m.c.hsu, “visual sentiment analysis on twitter data streams”, ieeesymposium on visual analytics science and technology, vol. 20, october 23-2014, providence, rhode island, usa.

[18]. M. Ghiassi, j. Skinner, and d. Zimbra, “twitter brand sentimentanalysis: a hybrid system using n-gram analysis and dynamic artificialneural network”, expert systems with applications, vol. 40, issue 16, pp 6266-6282, november 2013.