Education on Communal Control of Reduce in Country consuming Support Vector Machine

Mr Manoj Patel[#1]

[#] *Faculty of Computer Science & Applications,*

Gokul Global University, Sidhpur,

India

Abstract—Social networking plays a very important role to express public view on different politics activities, and organisation. Analysing and mining public reviews on demonetization is a complex task that involves understanding the various perspectives and opinions of people regarding this policy. Analysing the positive and negative reviews of people on the decision to transition India into a cashless or digital economy, as well as its impact on the economy, market, and exposure of black money, through Twitter data can provide valuable insights. Here's a more detailed step-by-step guide on how to conduct this analysis:

***Keywords*— Demonetization, currency, black money, corruption, cashless, digital economy, SVM, Machine learning**

I. INTRODUCTION

Demonetization is the process of ending the currency in the form of notes and coins as the legal tender of a country. In the Demonetization the old currency is retrieved or changed into the new form of currency. Indian Economy had a great impact of Demonetization At present, due to the use of internet people use different platforms to express their reviews. People express their opinion by messages on twitter, Facebook and google+ [1] . People share their opinion about religious and political issues on the microblogging websites. So, these microblogging sites are valuable source to know about public's opinion and sentiments. Sentiments analysis helps to know a big about public opinion about any decision or for taking a decision. So, let's see what people of India think about demonetization. There are many techniques to analysis the sentiments of people So in this paper we explain the machine learning classifiers such as maximum entropy, support vector machine (SVM) for finding the accuracy of sentiments [2].

Refer below Point.

1. Positive
2. Negative
3. Neutral

Positive Sentiments

- Positive Sentiments

Positive sentiments explain the positive views of the public by their messages and text. Happiness, joy and smiles are positive sentiments of the emotions. In case of political decision, the review shows that people are happy with the work or by the decision.

- Negative Sentiments

Negative sentiments explain the negative views of the public by their messages and text. Disappointment, jealous, hate and sadness are negative sentiments of the emotions. In case of political decision, the review shows that people are not happy with the work or by the decision

- Neutral Sentiments

Their opinion is neither preferred nor neglected. They didn't take interest in the decision.

II. SENTIMENTS CLASSIFICATION

In sentiments classification various methodologies are used to explain the sentiments. In the document level sentiment classification, the whole documents are regraded a information unit. There are three methods used for the classification are Lexicon based, machine learning based and rule-based method [1]

III. LEXICON BASED METHOD

Lexicon based method is used for analysis the sentiments of individual words and emotions.

In the lexicon there is an emotional dictionary is available for the text. Every word present in the text or message are compared to the emotion dictionary. Then the polarity value of the word is added to the total polarity value of the text or the message. If we found a positive match then the value of the score is added to the pool score of the word for the next input. For example; the word "Sweet" is matched with the word "sweet" of the dictionary which is interpreted as positive increase in polarity score of the blog. The positivity of the polar score classified that it is positive, otherwise it is negative. Subjective lexicon is a part of MPQA (manufacture product quality assessment) which contains the news articles from different source [3]. The subjectivity of lexicon follows the terms of the GNU General License.

In Lexicon Method First we calculated the weights of all training texts and the classified text Weight of the text will be defined as average weights of emotional words from dictionary presented. All texts of the sentiments are placed in unidimensional emotional space. To improve the accuracy of the classification of the text that were too close to another sentiment class text were excluded from consideration. The proportion of deletions of the text was determined by the cross-validation method. Then the average weights of training texts for each sentiment will be obtained.

The classified text of the sentiment was referred to the class which was located closer in the unidimensional emotional space.

IV. MACHINE LEARNING APPROACH

Machine learning approach uses ML algorithm and linguistic feature. Machine learning is used for sentiments analysis [5]. In the machine leaning two sets a training and a test set are required. A Training set in machine learning is used to differentiating characteristics of the documents and the test set is used to check the performance of the classifier. There are many techniques like Naïve Bayes (NB), maximum entropy (ME) and support vector machine (SVM) are used to classify the reviews[1]. For solving various problems of computational linguistics both methods Maximum Entropy and SVM use a vector modal of the text to obtain the vector modal for only one emotional dictionary.

It has been proven that in the term of accuracy and efficiency SVM is out performs as compare to other sentiments analysis tools used[8]. In this paper we explain the sentiments analysis techniques for the classification of extremist web forum posting in multiple languages by avail of stylistic and syntactic feature[1] . To improve the feature selection, they deduced new algorithm entropy weighted genetic algorithm(EWGA)which is hybrid genetic algorithm that use information gain heuristic [7].

Feature Selection in Sentiments Analysis

- Term frequency or Identifier frequency(TF-IDF)

The TF-IDF represents the consideration of the counting the words and the terms used. These words or terms may be uni-grams, bi-grams and high order n-grams. In this paper SMS spam collection are detected using machine learning o with TF-IDF is used for feature [9]

- Part of Speech tags(POS)

English is an ambiguous language so one word can have different meaning depending upon Its use. So, POS is used to disambiguate sense which is used for finding adjectives to guide feature selection because they are important opinion indicators [6].

- Syntax and negation

The use of syntactic features and collections are used to improve performance. In classification of short texts, algorithms using syntactic features with n-gram features were reported to yield the same performance. The negation of words can change the opinion orientation like not good is equivalent to bad.

PROPOSED APPROACH FOR SENTIMENTS ANALYSING

To propose a model for analysing sentiments of demonetization using support vector machine which explains polarity of twitter data and performances vector [1] .

The Proposed Methodology is shown below. Sentiment analysis on tweets is performed on 3rd stage. At 1st stage pre-processing is executed. After that Feature vector is created using applicable features.

A. Data Collection

We collected data from twitter related to demonetization. The data from twitter is gathered using Search API and Streaming API which is officially provided by Twitter. The Search API allows developers to search for a specific word or a phrase from tweets. But one of the restriction imposed by Twitter is that the Search API produces 1500 tweets at a time. Hence, to gather more tweets one has to use Streaming API which record tweets data in real time.

B. Data Pre-Processing

In data mining process the cleaning of data is a important process. As tweets contain several syntactic features which are not be useful for analysis that's why Cleaning of Twitter data is necessary. The pre-processing of the data is done in such a way that processed data is only in terms of words that can easily classify the class. To obtain processed tweet we create a code in Python.

- Language detection

We are interested in English text. So, all tweets have been separated into English and non- English data. So, for this detection we used NLTK's language detection feature for this.

- Removing punctuation and URL

Twitter has lot of unwanted punctuation, URLs, hash tag, @, duplicates which need to be removed to get processed and clean data for analysis. To make data more informative for the machine learning algorithms, a pre-processing method was implemented for eliminating them. The following table explains all the feature that have been removed.

- Tokenize

Tokenizes is a process which divide strings into lists of substrings also known as Tokens. Tokenized text is used to separate out the unnecessary symbols and punctuations and filters out those words that can be add value to the sentimental polarity score of the text.

- Stop words

In information retrieval common words such as is, am, are, in, of etc don't have much relevance, since their appearance in a post does not provide any useful information in classifying a document.

- Stemming

Stemming is a technique that put word variations like large, larger, largest all into one bucket, effectively decreasing entropy and increasing the relevance of the concept of large. In other words, stemming reduce token or words to their root form.

C. Classifier

We have used Support Vector machine (SVM) classifier in our approach. It is one of the popular and powerful techniques used for non-linear binary classification task. It optimizes the procedure of maximizing predictive accuracy while automatically avoiding over-fitting of the training data. SVM projects the data into a kernel space. There is a different kernel parameter used as a turning parameter to improve the classification accuracy [10]. Further, it builds a linear model in that kernel space. In a clear way, it maps the feature vector into high-dimensional feature space. In this model, we have used dot kernel function [1] The Dot Kernel is defined by

$$K(x, y) = x*y$$

It is inner product of x and y.

For classification process, each set was divided in two parts one for training and second for testing, by ratio 4:1, it means 4/5 parts were used for training and 1/5 for testing. Then training performed with 10 folds cross validation for classification.

Here we use two classifiers:

[1] Naïve Bayes Classifier

[2] Support Vector Machine

Naïve Bayes Classifier

Naïve Bayes Classifier is a easy approach to get the performance similar to the other techniques. Baseline Naïve Bayes classifier categorize the sentiments into 3 categories; positive, negative and neutral [20]

Support Vector Machine

Support Vector Machine is a supervised learning technique which is applicable for both classification and regression. The concept of SVM is on decision planes that defines the boundaries of decision.

SVM is mostly used for text classification task such as category assignment, detection spam and sentiment analysis. The function of SVM is given as;

$$H(x) = z^n \varphi(x) + c$$

Where x is represented as the feature vector. And z is the vector of different weight and the aid of Ø and c represents bias vector of the Non-linear mapping characteristic.
In SVM we use 3 labels to analyse the sentiments like 0,1 and 2. Here 0 stands for positive ,1 stands for negative and 2 stands for neutral. So, the tweets can be represented in the sequence of o's,1's and 2's. Now this feature vector and class labels are given to SVM classifier to classify the tweets in positive, negative and neutral[20].

D. Performance Parameter for Evaluations

For Evaluation of performance parameter, a confusion matrix is used by a classification system that contains information about actual and predicted classifications. The confusion matrix is also known as error matrix [24]. Performance of the systems is commonly evaluated by using the data in the matrix. The following table is a special kind of contingency table having 2 dimensions. A confusion matrix is formed from the four outcomes that produced as a result of binary classification.

- Accuracy

Accuracy (ACC) is calculated as the number of all the correct predictions divided by the total number of the data.

$$\text{ACCURACY} = \frac{TP + TN}{TP+TN+FP+FN}$$

- Precision

Precision is usually calculated as the number of correct positive predictions divided by the total number of positive predictions.

$$\text{PRECISION} = \frac{TP}{TP+FP}$$

- Sensitivity (Recall or True positive rate)

Sensitivity is defined as number of correct positive predictions divided by the total number of positives.

RESULT

On demonetization around 800 user's reviews achieved on online review data. Reviews from twitter were collected and all reviews were formatted according to the CSV files where review text and ID are only two attributes. Here we analyse the data set based on accuracy given by SVM. According to the survey the result obtained is positive that means people support Demonetisation. The obtained on the dataset has 63% accuracy and 61% is precision.

In confusion matrix:
True Negative (TN) =31, True Positive (TP) = 95, False Positive (FP) = 10, False Negative (FN) = 64, total =200. So, according to formula of accuracy, precision, classification error discussed above.

Accuracy calculated is 63% by dividing TP + TN to total number i.e. 200. Similarly, precision is calculated TP /predictive yes is 61%. The entire performance vector obtained is explained in Kappa is essentially a measure of how well the classifier performed is 274, Area Under Curve(AUC) is 780.

First of all we have to remove the Stop word from the tweets then we have to replace the words which are occurring twice in any particular word, with the two words that are trimming and repeated more than once[20] .

V. FUTURE WORK

In this survey we Analysis the review on twitter and after that recognized that how many positive, negative and neutral text are used. This demonetization will help us to eradicatethe black economy which is put in cash. Also, we moved towards the digital economy. After this evaluation and receiving user feedback, it is clear that the several improvements in the system are needed. The Sentiment Analysis need to incorporate negation and emphasis handling for improving its classification accuracy.

Future work includes to determine their features for the political decision in detail. We can only speculate future macroeconomic effects of demonetization.

V. CONCLUSION

Sentiment Analysis is one of the important research areas which helps us in summarizing opinion and reviews of public. In this research paper we analysis peoples' sentiment on demonetization. We analysed tweets on demonetization using SVM classifier. In this experiment we found that people have positive attitude for demonetization and they accept it as good decision for Indian politics. In our survey we obtained performance vector such as accuracy, precision, true positive rate. Calculated accuracy is 63.00% and precision obtained is 61%. Overall polarity obtained is positive.
So, we conclude that Machine Learning Technique is more efficient and easy as compared to the symbolic technique. These techniques can be easily adopted for the twitter data but sentiment analysis is actually very difficult to identify the emotional words from the tweets due to the presence of white spaces, misspellings and many other things. Now if we have to overcome these types of problems we can created feature vector.
Here we use the keyword demonetisation using hash tag and then we use the feature vector for extracting tweets. The features can be divided in two phases: First phase can be done by extraction of twitter with specific words from that removed the text now extracted feature vector is transformed into normal text and it will store in word cloud of a twitter then it can be done in graphical representation.

VI. REFERENCES

1. Uma Aggarwal, Gorav Aggarwal, "Sentiment Analysis on Demonetization using Machine Learning".
2. H. Tang, S. Tan, X. Cheng, "A survey on sentiment detection of reviews", Expert Systems with Applications, Vol.3 Issue no.7, pp.10760-10773,2009.
3. Derrick L. Cogburn, Fatima K. Es16_Apr_2018pinoza-Vasquez, "From networked nominee to networked nation: examining the impact of web 2.0 and social media on political participation and civic engagement in the 2008 Obama campaign", Journal of Political Marketing, Vol. 10, Issue. 1-2, pp. 189-213, 2011.

4. Anand Pandey, 12 December 2016,Sentiment Analysis on Demonetization https://bigishere.wordpress.com/2016/12/12/sentiment-analysis-on-demonetization-pig-use-case/.
5. Bo Pang, Lillian Lee, Shivkumar Vaithyanathan, "*Thumbs up?: sentiment classification using machine learning techniques*",
6. B. Pang and L. Lee, "Opinion mining and sentiment analysis".
7. A. Abbasi, H. Chen, and A. Salem, "Sentiment analysis in multiple languages: Feature selection for classification in web forums", In ACM Transactions on Information Systems, vol. 26, , pp. 1-34, 2008.
8. J. Kaur, S.S. Shera, S.K. Shera, "A Systematic Literature Review of Sentiment Analysis Techniques", International Journal of Computer Sciences and Engineering, Vol.5, Issue.4, pp.22-28, 2017.
9. G. Jain, B. Aggarwal, "Spam Detection on social Media Text", International Journal of Computer Sciences and Engineering, Vol.5, Issue.5, pp.63-70 2017.
10. S Parvathavardhini Manju, "Analysis on The Machine Learning Techniques", International Journal of Computer science.
11. *LIBSVM* — A library for support vector machines, available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm/.
12. Liu B. (2012), Sentiment analysis and opinion mining, Morgan & Claypool Publishers.
13. *Mystem,* available at: http://company.yandex.ru/technology/mystem.
14. Research of lexical approach and machine learning methods for sentiment analysis
15. *Pang B., Lee S.* Sentiment classification using machine learning techniques, Proceeding theEmpirical Methods in Natural Language Processing
16. *Pang B., Lee L.* (2008), Opinion Mining and Sentiment Analysis, Foundations and TrendsR in Information Retrieval
17. *Saif H., He Y., Alani H.* (2012), Alleviating data sparsity for twitter sentiment analysis
18. Sarvabhotla K, Sentiment classification: lexical similarity on based approach for extracting .
19. Wilson T, Hoffmann P Recognizing contextual polarity in phrase-level sentiment analysis. Empirical methods in natural language processing. Association for Computational Linguistics, .
20. Trends and Applications in Knowledge Discovery and Data Mining, Heidelberg, Germany.
21. Kameswara Rao and Tarun Nara" Using Supervised and Machine Learning Techniques for Improving the Accuracy of Opinion Mining on Tweets" .
22. Chesley P, Vincent Using verbs and adjectives to classify blog sentiment
23. Twitter Sentiment Analysis Of Movie Reviews Using Machine Learning Techniques Akshaya