IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-5, Issue-6 E-ISSN: 2347-2693

Sentiment Analysis on Demonetization using SVM

Uma Aggarwal[1*], Gaurav Aggarwal [2]

[1]Department of Computer Science, Jagannath University, Bahadurgar, India
[2]Department of Computer Science, Jagannath University, NCR, Haryana, India

[*]*Corresponding Author: er.umaaggarwal@gmail.com*

Available online at: www.ijcseonline.org

Received: 11/May/2017, Revised: 14/May/2017, Accepted: 10/Jun/2017, Published: 30/Jun/2017

***Abstract*—** Sentiment Analysis is an area of interest over the last decade. The social networking is one of the important sources for users to know express the views on different organizations, product, and politics. In this work, we focus on mining sentiments and analyzing public review on demonetization. Demonetization was one of the biggest political decisions taken in year 2016 which affected each and every person in India. In demonetization 500 and 1000 rupees currency was banned more over a new 2000 rupees note was introduced in currency. This affected economy, market and exposed black money also. We worked on twitter data for demonetization. It aims to analyzing positive and negative of tweets reviews as sentiment classification task. The raw dataset collected is preprocessed by cleaning unwanted text, tokenized and used for polarity classification of data corpus.

***Keywords*—** Sentiment Analysis , Classification, SVM, Machine learning

I. INTRODUCTION

With increase in use of internet, people use different platforms such as microblogging, forums and social networks to express their views. These applications help users to express their views, emotions and sentiments towards products, people and life in general. Lot of messages are done daily on popular web-sites that provide services for microbloging such as Twitter, Facebook, Google+. Users express their feeling about their life, share opinions on variety of topics and discuss current issues. Lot of users' post their views about products and services they use and express their political and religious views, hence, microbloging websites has become valuable sources of peoples' opinions and sentiments. Sentiment analysis helps a lot in politics as it helps to analysis public opinion on a particular decision and used in taking decision. Also, it helps in analysis election pools though publics' reviews. So, demonetization was of one the biggest political decision of 2016 in India which deeply affected everyone. So, we have used twitter data on demonetization a for sentiment analysis. So see what people think about demonetization. Different ways are used for analysis sentiment technology. This paper explain machine learning classifiers such as naive Bayes, maximum entropy and support vector machine (SVM) are used in for sentiment classification to achieve accuracies that range from 75% to 83%, in comparison to a 90% accuracy or higher in topic based categorization[1].

There are three classes of sentiments .i.e. positive, negative and neutral sentiments on basis of which we analysis data.

- **Positive Sentiments**

It explains positive view of the speaker about the text. Emotions with positive sentiments shows happiness, joy, smile etc. In case of political reviews, positive reviews about political decision or a politician shows people are happy with their work.

- **Negative Sentiments**

Negative sentiment shows negative attitude of the speaker about the text. Emotions with negative sentiments shows disappointment, sadness, jealousy, hate etc. In case of political reviews, negative reviews about the politician or political decision are more shows people are not happy with his work.

- **Neutral Sentiments**

Neutral sentiment shows neutral views about the text. Text is neither preferred nor neglected.

Rest of the paper is organized as follows, Section I contains the introduction of this paper. Section 2 explains classification sentiment analysis. Section 3 is proposed approach we used to analysis sentiment on demonetization. Section 4 contains result of this analysis. Section 5 is conclusion and future work. Section 6 is references

II. SENTIMNET CLASSIFICATION

This section explains about various methodologies used in sentiment classification. Mostly, the task refers to document-level sentiment classification in which whole document is regarded as an information unit. There are three type of

methods used in classification are: lexicon-based, machine learning-based and rule-based methods.

A. Lexicon-Based Methods

Lexicon is very important method in sentiment analysis. A lexical approach avail a dictionary or lexicon of pre-tagged words therefore, it provides sentiment information about the smallest linguistic unit. Each word present in a text is compared with word present in the dictionary. Every encountered word is matched with that dictionary. Then its polarity value is added to the total polarity score of the text. If we found a positive match, then score is added to the total pool of score for the input text. For example, if a word "beautiful" is matched with the word "beautiful" of dictionary, which is interpreted as positive in dictionary, then the total polarity score of the blog is increased. If the total polarity score of a text is positive, then that text is classified as positive, otherwise it is classified as negative. MPQA opinion contains news articles from different news sources. Subjective lexicon is part of MPQA Opinion Corpus[2]. The subjectivity lexicon is made available under the terms of GNU General License .

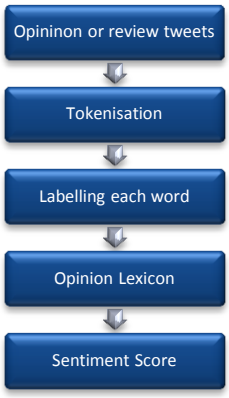

Fig 1 generic architecture of lexicon approach

Machine Learning Approach

The Machine Learning Approach (ML) uses the famous ML algorithms and uses linguistic features. The first work on machine learning for sentiment analysis is [3]. In a machine learning based techniques, two sets of documents are required that are training and a test set. A training set is used as human automatic classifier to learn the differentiating characteristics of documents whereas a test set is used to check how well the classifier performs. Machine learning tasks are typically classified into two broad categories accordingly approaches are applied. There is different machine learning techniques have been used to classify the reviews. Various Machine learning techniques are like Naive Bayes (NB), maximum entropy (ME), and support vector machines (SVM) that are used to achieve great success in sentiment analysis. It has been proved in the recent research that SVM outperforms than the sentiment analysis tools used and other techniques in terms of accuracy and efficiency [6].

This paper explains sentiment analysis techniques for classification of extremist web forum postings in multiple languages (English and Arabic) by avail of stylistic and syntactic features. They deduced new algorithm entropy weighted genetic algorithm (EWGA) which is hybrid genetic algorithm that uses the information gain heuristic to improve feature selection [5].

The following are features used in sentiment analysis.

- TF-IDF (Term frequency or Identifier frequency)

The TF-IDF represents counting the terms or words being taken into consideration. These terms or words can be uni-grams, bi-grams or higher order n-grams. In research paper, SMS Spam collection are detected using machine learning in which Bag Of Words with TF-IDF is used for feature selection[7].

- POS -Part Of Speech tags

English is an ambiguous language by nature, so one particular word can have different meaning depending upon its context of use. So, POS is used to disambiguate sense which is used to guide feature selection [4].

- Syntax and negation

The use of collocations and other syntactic features can be used to improve performance. In classification of short texts, algorithms using syntactic features and algorithms using n-gram features were reported to yield the same performance.

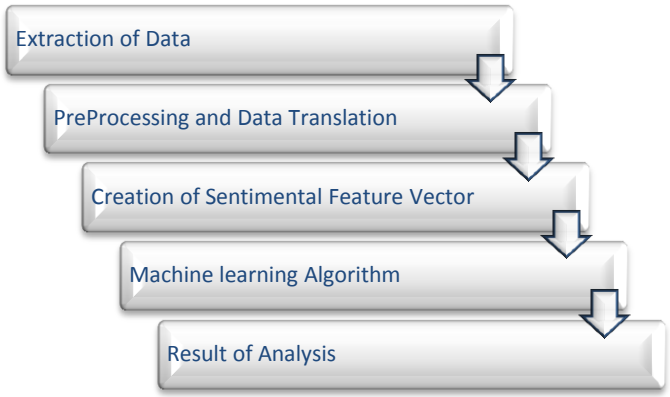

Fig 2 Generic architecture of machine learning

III. PROPOSED APPROACH FOR SENTIMENT ANALYSIS

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

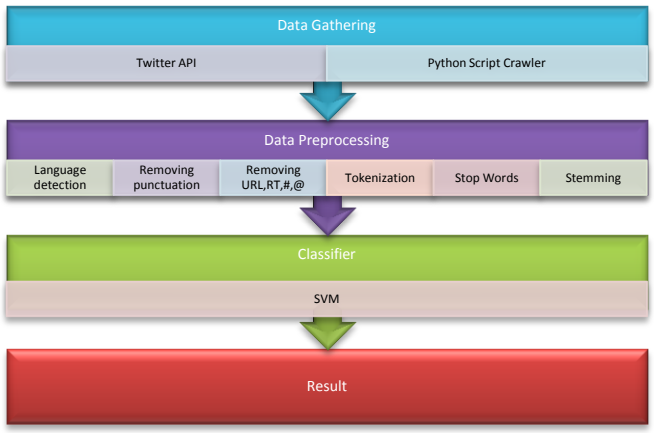

Fig 3 Flowchat of proposed approach

A. Data Collection

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

We used the python language to carry out our experiment. Tweepy is one of the open sources Python library which helps python to communicate with twitter and use its API to collect data so that we can use it in our program.

B. Data Preprocessing

Data cleaning is an important part of the data mining process. Cleaning of Twitter data is necessary as tweets contain several syntactic features that are not be useful for analysis. The pre-processing is done in such a way that processed data is only in terms of words that can easily classify the class. We create a code in Python to obtain processed tweet. This code is used to achieve the following functions:

- *Language detection*

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

- *Removing punctuation and URL*

There is 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 are removed.

Features to be removed	Full form	Explanation
WWW. / https /	URL	Typically a link
@	Mention	Lag to mention another user
#	Hash Tag	Used to tag a tweet
Yahoooooo	Letter Duplicity	Sign of excessive joy or sad
My name IS	Words in different cases	Used to write something
? ' " ; . < >	Punctuation	Used for special purpose
India I s	Additional White spaces	Misspelling or slang word
RT	Retweets	Reposting another's tweet

TABLE 1
Features to be removed in pre-processing of text

- *Tokenize*

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

Example

Plain text- "Demonetizations is good for Indian politics"

Tokenized - Demonetizations, is, good, for, Indian, politics

- *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 techniques 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

Sentiment analysis is used to classify text as positive, negative. We have used Support Vector machine (SVM) classifier in our approach. It is one of the popular and powerful techniques for non-linear binary classification task. It optimizes procedure of maximizing predictive accuracy while automatically avoiding over-fitting the training data. SVM projects the data into a kernel space. There is different kernel parameter used as a turning parameter to improve the classification accuracy [8]. Further, it builds a linear model in that kernel space. In a clever way, it maps feature vector into high-dimensional feature space. In this model, we have used dot kernel function. The Dot Kernel is defined by

$$k(x,y) = x*y \quad (1)$$

it is inner product of x and y.

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

D. Performance Parameter for Evaluations

For Evaluation of performance parameter a confusion matrix is used that contains information about actual and predicted classifications done by a classification system. A confusion matrix is also known as an error matrix [9]. Performance of systems is commonly evaluated using the data in the matrix. The following table is a special kind of contingency table, with 2 dimensions. A confusion matrix is formed from the four outcomes produced as a result of binary classification.

- *Accuracy*

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

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP+TN+FP+ FN}} \quad (2)$$

- *Precision*

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

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (3)$$

- *Sensitivity (Recall or True positive rate)*

  Sensitivity is calculated as the number of correct positive predictions divided by the total number of positives. It is also called recall (REC) or true positive rate (TPR).

$$\text{SN} = \frac{\text{TP}}{\text{TP+FN}} \quad (4)$$

IV. RESULT

The online review dataset consists of around 800 user's review archived on demonetization. And for, Twitter tweets review were collected and each review were formatted according to csv file where review text and id are only two attributes. Here, we analyze the dataset based on accuracy given by SVM. According to our, experiment result obtained is positive that means people support demonetization. The result obtained on dataset has accuracy around 63% and precision is around 61%. In our experiment, obtained confusion matrix is explained in table 2:

	Predicted Negative	Predicted positve
Actual negative	31	10
Actual Positive	64	95

Table 2 : Confusion matrix obtained in experiment

In obtained 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 figure 4. Kappa is essentially a measure of how well the classifier performed is .274 , Area Under Curve(AUC) is .780

PerformanceVector

```
PerformanceVector:
accuracy: 63.00% +/- 10.05% (mikro: 63.00%)
ConfusionMatrix:
True:   negative        positive
negative:       31      10
positive:       64      95
classification_error: 37.00% +/- 10.05% (mikro: 37.00%)
ConfusionMatrix:
True:   negative        positive
negative:       31      10
positive:       64      95
kappa: 0.274 +/- 0.165 (mikro: 0.238)
ConfusionMatrix:
True:   negative        positive
negative:       31      10
positive:       64      95
AUC (optimistic): 0.780 +/- 0.087 (mikro: 0.780) (positive class: positive)
AUC: 0.780 +/- 0.087 (mikro: 0.780) (positive class: positive)
AUC (pessimistic): 0.780 +/- 0.087 (mikro: 0.780) (positive class: positive)
precision: 61.91% +/- 15.56% (mikro: 59.75%) (positive class: positive)
ConfusionMatrix:
True:   negative        positive
negative:       31      10
positive:       64      95
false_positive: 6.400 +/- 3.007 (mikro: 64.000) (positive class: positive)
ConfusionMatrix:
True:   negative        positive
negative:       31      10
positive:       64      95
```

Figure 4 : performance Vector calculated in experiment

V. CONCLUSION

Sentiment Analysis is one of the important research areas as which helps 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 people have positive attitude for demonetization and accept it as good decision for Indian politics. In our experiment 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.

Future work include to determine their features for the political decision in detail i.e. make polarity check on different features such as total no of hashtag , combining two approaches etc.

VI. REFERENCES

[1] H. Tang, S. Tan, X. Cheng, "*A survey on sentiment detection of reviews*", Expert Systems with Applications, Vol.36 , Issue no.7, pp.10760- 10773,2009.

[2] Derrick L. Cogburn ,Fatima K. Espinoza-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.

[3] Bo Pang, Lillian Lee, Shivkumar Vaithyanathan, "*Thumbs up?: sentiment classification using machine learning techniques*", in

Proceeding ACL Conference on Empirical Methods in Natural Language Process., vol. 10, pp. 79-86, Philadelphia, PA, 2002.

[4] B. Pang and L. Lee, "*Opinion mining and sentiment analysis*", Foundations and Trends in Information Retrieval Vol no. 2, Issue.1-2, pp. 1–135,2008.

[5] A. Abbasi, H. Chen, and A. Salem, "*Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums*", In ACM Transactions on Information Systems, vol. 26, Issue 3, pp. 1-34, 2008.

[6] J. Kaur, S.S. Sehra, S.K. Sehra, "*A Systematic Literature Review of Sentiment Analysis Techniques*", International Journal of Computer Sciences and Engineering, Vol.5, Issue.4, pp.22-28, 2017.

[7] 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.

[8] S . Parvathavardhini and S . Manju, "*Analysis on Machine Learning Techniques*", International Journal of Computer Sciences and Engineering, Vol.4, Issue.8, pp.59-77, 2016.

[9] Stehman, Stephen V, "*Selecting and Interpreting measures of thematic classification accuracy*", Remote Sensing of Environment , Vol .62, Issue .1 , pp . 77-89, 1997

Authors Profile

Uma Aggarwal pursed Bachelor of technology from banasthali University, Banasthali in 2015 and is currently pursuing Master of technology from Jagannath University.

Dr. Gaurav Aggarwal received M.Tech (Computer Science and Engineering) degrees from Maharshi Dayanand University in 2008. Presently, he is working as an Assistant Professor in Computer Science and Engineering Department in Jagannath University, NCR,Haryana,India. Her areas of interest are software Reliability and Neural Networks.

