International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 6, Issue 5, September- October 2017 ISSN 2278-6856

# Sentiment Analysis of Social Media Data

**K.SandhyaRani[1], Dr. D. Kavitha[2]**

[1]PG Student, Computer Science and Engineering Dept, GPREC,
Kurnool (District), Andhra Pradesh-518007, INDIA.

[2]Professor, Computer Science and Engineering Dept, GPREC,
Kurnool (District), Andhra Pradesh-518007, INDIA.

***ABSTRACT: Social media networks have received wide attention today. Now a days they have become the main origin of user generated content. Analysis of this user generated content resulted wide applications like Sentiment Analysis(SA), Event detection, Forecasting systems, Topic modeling, Crowd-sourcing, Spam detection etc. In this project tweets on "Demonetization" have been collected using Twitter Streaming Application Programming Interface (API). Naive Bayes, Maximum Entropy Classifiers are used to analyze the tweets. ON comparing the results we observed that Maximum Entropy classifier gives better results than Naive bayes classifier.***

**Keywords:** R programming, data analytics, SA, Demonetization in India

## 1. INTRODUCTION

Sentiment Analysis (SA) is a kind of data mining that computes the inclination of users opinions by Natural Language Processing (NLP), Text Analysis and computational linguistics to capture the sentiment content words. SA intends to verify the outlook of a writer or a speaker (through some social media posts or tweets) with respect to a specific topic of a document. Globally, business enterprises can leverage opinion polarity and sentiment topic recognition to gain deeper understanding of the users and the overall scope. Consequently, these insights can advance competitive intelligence and develop the service, thereby constructing a better brand image and providing a competitive edge.

SA can be deemed as a sorting process. There are three levels in SA: Document-level, Sentence-level, and Aspect-level SA. Document-level SA intends to categorize an opinion document as expressing a positive or negative opinion and deems the whole document as a basic information unit (talking about one topic). SA aspires to categorize sentiment expressed in each sentence. The initial step is to know whether the sentence is subjective or objective [1].

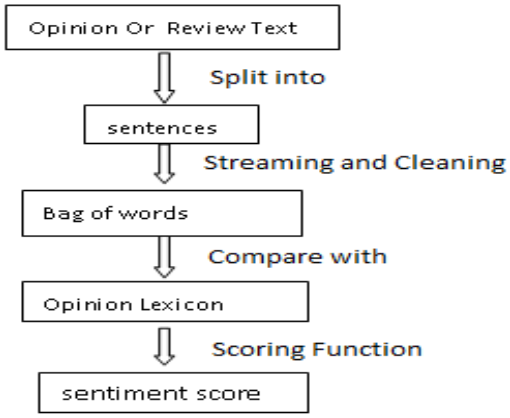



**Fig.1:** Sentiment Analysis procedure on product reviews

## 2. RELATED WORK

Jie Yang et al.[3] has proposed a Douban Learning framework for analyzing movie reviews from the User Generated Content. He introduced three modules: 1. Data crawler, which collects raw samples, 2. Feature generation, to identify subjective information, 3. Content mining, to extract high level features and concluded that the framework provides flexible, efficient for analyzing huge amount of social media data.

Desheng Dash et al.[4] has proposed Decision support approach which was developed using Support Vector Machine (SVM), Generalized autoregressive conditional heteroskedasticity (GARCH) modeling and Sentiment Analysis. He collected data from an online website called Sina Finance and found it helpful in determining the decision of investors and also in predicting future stock analysis. On comparing the results, observed that static machine learning approach has higher accuracy than semantic approach.

Xialion Zheng et al.[5] has studied Appraisal Expression Pattern (AEP) approach for determining aspect and sentiment words. He proposed AEP-based Latent Dirichlet Allocation model (AEP-LDA), which is a sentence-level, probabilistic generative model that assumes all words in a sentence are extracted from one topic. AEP-LDA model

performs better than AEP approach in identifying aspect and sentiment words.

Sven Rill et al. [6] designed PoliTwi method to determine emerging topics in twitter. About 40,00,000 tweets are collected before and during the parliament election 2013 in Germany for analysis and observed the results that new topics emerge earlier in Twitter when compared to Google trends.

Thien Hai Nguyen et al. [7] proposed Topic centric model for predicting the stock price movement using the sentiments from online blogs. The data is extracted in two ways one is Joint sentiment/Topic (JST) , to extracts sentiments and topics and other is Aspect based sentiment, to identify sentiments and topics. On comparing the accuracy of 18 stocks in one year transaction the proposed model achieved 2.07% improved performance .

Duyu Tang et al. [8] proposed Joint segmentation and Classification framework for sentence level classification.. He described three models:

1. Candidate generation model, generates segmentation candidates for each sentence
2. Segmentation ranking model, assigns score to each sentence
3. Sentiment classification model, predicts the sentiment polarity. The experiment results show that the proposed method performs comparably with state-of-the-art methods.

Farhan Hassan Khan et al.[9] presented A sentiment dictionary, SentiMI to build upon the mutual information evaluated from SentiWordNet 3.0 using its synsets. The SentiWordNet is used as label corpus to construct the SentiMI, which detects the sentiment polarity. Computing the mutual information, observed that there is improved performance of 7% in accuracy, 14% in specificity, 8 % in F-measure.

## 3. PROPOSED METHOD

In this paper, we have considered Twitter online networking as a data source for extracting tweets. By using R Programming language we performed SA on Demonetization Fig: 3.1 shows the process to determine the SA for tweets data on Demonetization. We have fetched tweets from various twitter users with different types of opinions and performed the SA using Naive Bayes and Maximum Entropy Algorithms

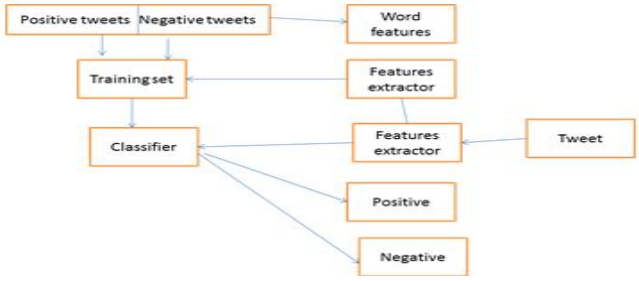



**Fig: 3.1** Procedure of Twitter data SA

**R Packages Used:**

In this paper, we have used below Packages of R programming for our research on Demonetization

library(twitteR)

library(tm)

library(sentiment)

library(snowballc)

library(wordcloud)

library(Rcurl)

library(ROAuth)

library(stringr)

library(maxent)

A. **Twitter Authentication and fetching the Tweets**

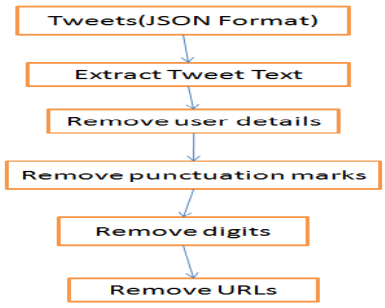
R language uses twitteR package to extract information from Twitter for Text Mining purpose. In order to get the connection between R console and Twitter , we need to set up a secure connection with Twitter using below keys.

```
consumer_key<- enter generated consumer key
consumer_secret<- enter generated consumer key
access_token<-enter generated access token key
access_token_secret<- enter generated access token secret key
```


**Fig: 3.2** Tweets fetching process

**B. Data Preprocessing**

Twitter API sends the tweets in Java Script Object Notation(JSON) format through Twitter REST API, so we are using few methods to cleanse the data for SA.We used the below process to extract the tweet from the json object for the sentiment analysis.




Below is the sample json (say ***xdata***) received from twitter API on demonetization.


```
{
    "text": "RT @AAPGujarat: News in Print: AAP protests in Gujarat
against #demonetization debacle (6/6) https://t.co/gikDcYQCqX",
    "truncated": true,
    "in_reply_to_user_id": null,
    "in_reply_to_status_id": null,
    "favorited": false,
    "source": "<a href=\"http://twitter.com/\" rel=\"no
follow\">Twitter for iPhone</a>",
    "in_reply_to_screen_name": null,
    "in_reply_to_status_id_str": null,
    "id_str": "546918022839009 28",
    "entities": {
"user_mentions": [
        {
"indices": [
                3,
                19
            ],
            "screen_name": "PostGradProblem",
            "id_str": "271572434",
            "name": "PostGradProblems",
            "id": 271572434
        }
```


Step 1: Extracting the tweet text from json object (xdata)

> "@AAPGujarat: News in Print: AAP protests in Gujarat against #demonetization debacle (6/6) https://t.co/gikDcYQCqX"

Step 2: Removing the user details

> "News in Print: AAP protests in Gujarat against #demonetization debacle (6/6) https://t.co/gikDcYQCqX"

Step 3: Removing the punctuation marks

> News in Print AAP protests in Gujarat against #demonetization debacle 66httpst.cogikDcYQCqX

Step 4: Removing the digits

> News in Print AAP protests in Gujarat against #demonetization debacle httpst.cogikDcYQCqX

Step5: Removing the URLs

> News in Print AAP protests in Gujarat against #demonetization debacle

## 4. ALGORITHMS

**Naive Bayes Classifier:**

The Naive Bayes (NB) classifier is the easy and most usually used classifier. In view of the appropriation of the words this model figures the back probability of a class. Bayes Theorem is utilized to foresee the probability that a given list of capabilities has a place with a particular label.

$$P(label|features) = \frac{P(label) * P(features|label)}{P(features)}$$

P(label) is the probability that is random feature set to the label. P(features | label) is the earlier probability that a given list of capabilities is being delegated a mark. P(features) is the prior probability that a given feature set is occurred.

**Table:** Sample Tweets and their features

| Tweets | Features |
|---|---|
| Happy to see another unblemished leader like@NitishKumar backing #DeMonetizationCONmen, #AAPholes, Commie & Socialist burning up.. CooL! | Happy, backing, burning, cool |
| #demonetization 's irony... most of the people don't even realize what they gaining but happy about what's someone else is losing. Envy!!!!!! | Gaining, Happy, losing |
| Modi #demonetizes and Hitler is Angry! @narendramodi @PMOIndia #demonetization https://t.co/54B7e9cRfv | Angry |
| RT @trending freaks: Did Common man angry on PM NarendraModi Demonetization Decision? MUST WATCH #pm modi #modi #demonetization... https://t… | Angry |
| @sardesairajdeep Thus spake an incompetent journalist who could not predict #Demonetization with all his network and informers...disgusting. | Incompetent, disgusting |
| Yes @yoginisd I think it's good the marriage broke, dowry marriages make the woman unhappy. Media is linking everything to #demonetization | Good, unhappy |

**Table:** Tweets with their features and Labels

| Tweets | Features | Label |
|---|---|---|
| Tweet1 | Happy | Positive |
| Tweet1 | Backing | Positive |
| Tweet1 | Burning | Negative |
| Tweet1 | Cool | Positive |

| | | |
|---|---|---|
| Tweet2 | Gaining | Positive |
| Tweet2 | Happy | Positive |
| Tweet2 | Losing | Negative |
| Tweet3 | Angry | Negative |
| Tweet4 | Angry | Negative |
| Tweet5 | Incompetent | Negative |
| Tweet5 | Disgusting | Negative |
| Tweet6 | Good | Positive |
| Tweet6 | Unhappy | Negative |

from above two tables, we can calculate the likelihood probability of the labels as below.

P(Positive|Happy)=P(positive)*P(Happy|positive)*P(backing|positive)*P(cool|positive)*P(gaining|positive)*P(Good|positive)*1/P(Happy)

i.e.*P(Positive|Happy)=0.46*1*1*1*1*1/0.15=1/3=0.33*

P(Negative|unhappy)=P(Negative)*P(burning|negative)*P(losing|negative)*P(angry|negative)*P(Incompetent|negative)*P(disgusting|negative)*P(unhappy|negative)*1/P(unhappy)

i.e. *P(Negative|Unhappy)=0.53*1*1/1=0.53*

and based on above calculations, *the unknown is 1-(0.33+0.53)=0.17*

Conclusion of Naive states that all highlights are autonomous, the condition could be reworked as follows

$$P(label|features) = \frac{P(label) * P(f1|label) * \ldots .. * P(fn|label)}{P(features)}$$

An improved NB classifier was proposed by Kang and Yoo [10] to take care of the issue of the propensity for the positive order precision to show up to roughly 10% higher than the negative . This creates a problem of decreasing the average accuracy when the accuracies of the two classes are expressed as an average value.

So the Naive Bayes classification says that the polarity levels of Indian Citizens on Demonetization as Positive 60%, Negatives 30% and neutral 10% and the results shows that majority peoples are reacting positively for the Demonetization.

**Maximum Entropy classifier :**

The Maximum Entropy (ME) Classifier (known as a restrictive exponential classifier) utilizes encoding to change over marked featureset to vectors. This encoded vector is then used to ascertain weights for each component that can be joined to decide the probability for a list of capabilities. This classifier is parameterized by an arrangement of X{weights}, which is utilized to consolidate the joint highlights that are created from a list of capabilities by a X{encoding}. Specifically, the encoding maps each C{(featureset, label)} combine to a vector. The probability of each name is then evaluated using the accompanying condition:

$$P(fs|label) = \frac{dotprod(weights, encode(fs, label))}{sum(dotprod(weights, encode(fs, l))forlinlabels)}$$

Labels: {Positive, Negative, Neutral}

Featureset (fs):{Happy, backing, burning, cool, gaining}

*P(Positive) + P(Negative) + P(Neutral) = 1*

Suppose that if the word “happy” appears in the tweets, then

*P(Positive|Happy)=0.5* and

then the distribution is adjusted as

*P(Negative|Happy)=0.25*

*P(Neutral|Happy)=0.25*

Encode(fs,label):

| Feature short form used in tweets | Encode |
|---|---|
| Awsm | Awesome |
| Hpy | Happy |
| Happyyyy | Happy |

Taking an example tweet:

Happy to see another unblemished leader like@NitishKumar backing #DeMonetizationCONmen, #AAPholes, Commie & Socialist burning up.. CooL!

Weights for the above tweet:

| | Happy | To | see | another | Unblemished | leader |
|---|---|---|---|---|---|---|
| Negative | 0 | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 |
| Positive | 1 | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 |
| Neutral | 0 | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 |

P(Positive|Happy)=prod(weights(posit),encode(fs,label))

i.e. *P(Positive|Happy)=1*1/6+1/6*1/6=0.19*

P(Negative|Happy)=prod(weights(neg),encode(fs,label))

i.e.*P(Negative|Happy)=0*1/6+1/6*1/6=1/36=0.027*

P(Neutral|Happy)=prod(weights(neu),encode(fs,label))

i.e

*.P(Neutral|Happy)=0*1/6+1/6*1/6=1/36=0.027*

Kaufmann [11] utilized ME classifier to distinguish parallel sentences between any dialect sets with little measures of preparing information. Alternate apparatuses that were created to consequently separate parallel information from non-parallel corpora utilize dialect particular systems or require a lot of preparing information. Their outcomes demonstrated that ME classifiers can create helpful outcomes for any dialect combine. This can permit the making of parallel corpora for some new dialects.

**Pseudo Code:**

In this paper, we have implemented the SA for finding the Indian Citizens opinion on Demonetization using the below steps.

Step1. Loading all the required packages into R

Step2. Twitter Authentication using oauth

Step3. Fetching tweets

Step4. Data preprocessing

Step5. Identifying the features to be used for the sentiments analysis

Step6. Implementing Naive Bayes Classification to classify emotion and polarity

Step7. Implementing Maximum Entropy classification to classify emotion and polarity

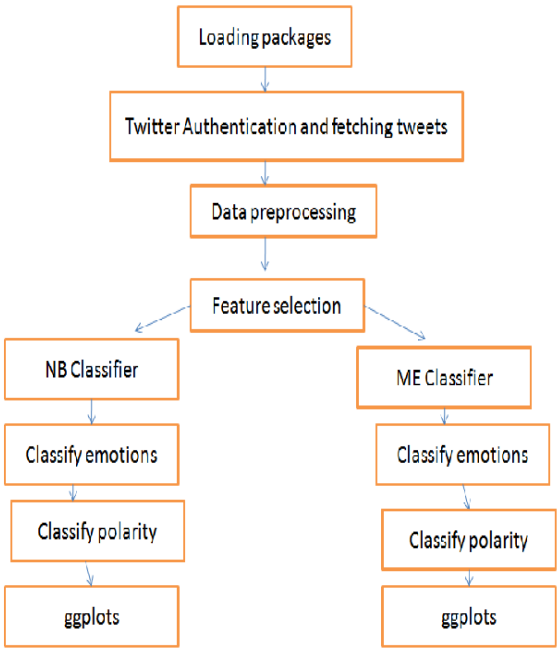



**Fig: 4.1** Heirarchy of SA on Naive Bayes and Maximum Entropy Algorithms

## 5. RESULTS AND DISCUSSION

In this paper, we observed the views of different users on the Demonetization. By analyzing the tweets

Naive Bayes classification shows the emotions of Indian Citizens on Demonetization are 30% as joy, 10% as sad and few are fear and disgust. Here Naive Bayes classification uses the predefined features as shown in below figure

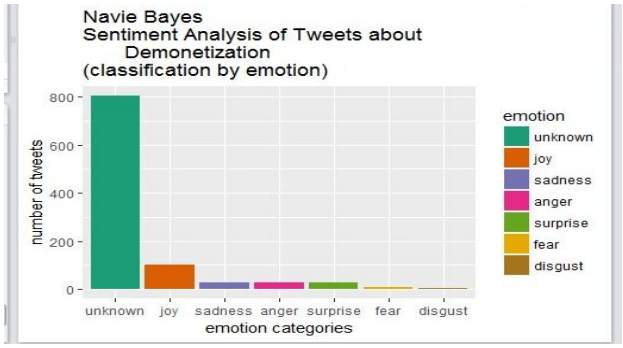



**Fig 5.1** Naive Bayes Classification by Emotion

Fig. 5.2 shows the polarity levels of users on Demonetization and demonstrates that majority of people are reacting positively.

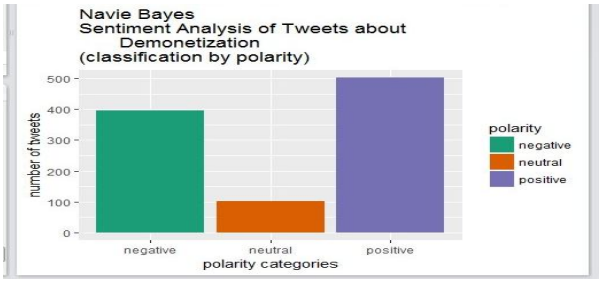



**Fig 5.2** Naive Bayes Classification by Polarity

Fig.5.3 shows the word cloud of Naive Bayes classification in which few opinions are not displayed due to its limited scope on features.

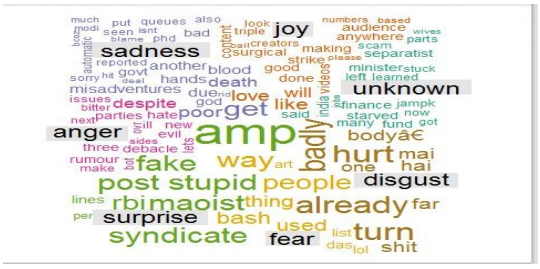



**Fig 5.3** Naïve Bayes Classification words cloud

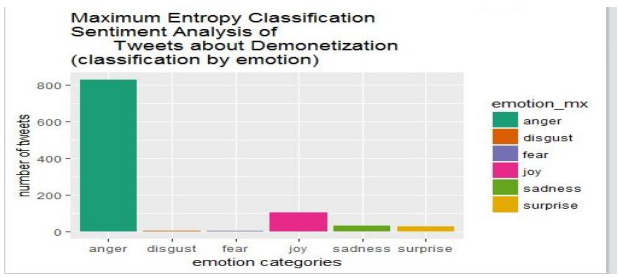



**Fig 5.4** Maximum Entropy Classification by Emotion

Maximum Entropy Classification shows the emotions of Indian Citizens on Demonetization are 70% as angry, 10% as joy and few are fear and disgust. Here Maximum Entropy classification uses the most repeated words to find the features in the tweets

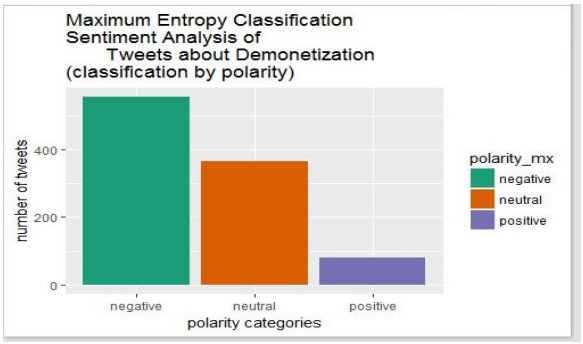



**Fig 5.5** Maximum Entropy by Polarity

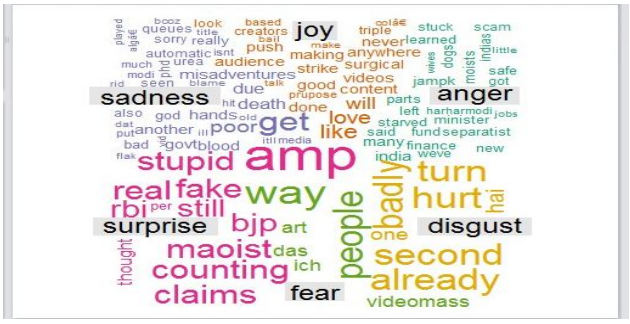



**Fig 5.6** Maximum Entropy by word cloud

## 6. CONCLUSION

In this paper, we analyzed tweets from Twitter on Demonetization which determines positive ,negative and neutral words. We performed text classification by using :Naive Bayes that also considers unknown words and determined 30% as joy,10%as sad,5%as disgust. Maximum Entropy captures only emotions and shows 70% anger 10%as joy and few are fear, disgust. On comparing two algorithms maximum Entropy gives best results than Naive Bayes.

## References

[1] Steven L. TextBlob. [Online].https://textBlob.readthe docs.org/en/dev.[cited Oct 20,2015]

[2] The statistics portal https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[3] Jie Yang, Brain Yecies Mining Chinese social media UGC:a big data framework for analyzing Douban movie reviews

[4] Desheng Dash Wu, Lijuan Zheng, and David L. Olson, A Decision Support Approach for Online Stock Forum SA, 2168-2216 2014 IEEE

[5] Xiaolin Zheng, Zhen Lin, Xiaowei Wang, Kwei-Jay Lin, Meina Song, Incorporating appraisal expression patterns into topic modeling for aspect and sentiment word identification, Knowledge-Based Systems 61 (2014) 29–47

[6] Sven Rill, Dirk Reinel, Jörg Scheidt, Roberto V. Zicari, PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level SA2014 Elsevier B.V.

[7] Thien Hai Nguyen, Kiyoaki Shirai, Julien Velcin, SA on social media for stock movement prediction (2015)

[8] Duyu Tang, Bing Qin, Furu Wei, Li Dong, Ting Liu, and Ming Zhou, A Joint Segmentation and Classification Framework for Sentence Level Sentiment Classification, 1750 IEEE, VOL. 23,20

[9] Farhan Hassan Khan, Usman Qamar, SentiMI: Introducing Point-wise Mutual Information with SentiWordNet to Improve Sentiment Polarity Detection, Applied Soft Computing · November 2015

[10] Hanhoon Kang, SeongJoonYoo, Dongil Han, Senti-lexiconand improved Naïve Bayes algorithms for SA of restaurant reviews, Expert systAppl 2012;39:9166-80

[11] Moraes Rodrigo, ValiatiJoa˜o Francisco, Gavia˜oNeto Wilson P. Document-level sentiment classification: an empirical comparison between SVM and ANN. Expert SystAppl 2013;40: 621–33

K.Sandhyarani received the Bachelor's. Degree in Information Technology from G.Pullaiah College of Engineering and Technology , Kurnool, A.P. Her research interests include Big data analytics.

Dr.D.Kavitha obtained her BTech degree from Sri Krishna Devaraya University, Anantapur in the year 2001and 2005 respectively she obtained Ph.D in Computer Science from Sri Krishna Devaraya University, Anantapur in 2012. she has presented thirty research papers in various National and International journals so far. Her research areas include Computer Networks and Network Security, Big data Analytics.